National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, tests publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c.-C-30.

NL-339 (r. 88/04)

Canada

THE UNIVERSITY OF ALBERTA

A SURVEY OF SCALING METHODS FOR THE MEASUREMENT OF ATTITUDES: A COMPARISON OF AMERICAN AND WEST-GERMAN ATTITUDES TOWARD ABORTION

by

ELISABETH MARIA TEN VERGERT

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF SOCIOLOGY

EDMONTON, ALBERTA

SPRING, 1988

Permission has been granted to the National Library of Canada to microfilm this thesis and to lend or sell copies of the film.

The author (copyright owner) has reserved other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without his/her written permission.

L'autorisation a été accordée à la Bibliothèque nationale du Canada de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur (titulaire du droit d'auteur) se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation écrite.

ISBN 0-315-42965-8

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR ELISABETH MARIA TEN VERGERT

TITLE OF THESIS A SURVEY OF SCALING METHODS FOR THE MEASUREMENT OF ATTITUDES: A COMPARISON OF AMERICAN AND WEST-GERMAN ATTITUDES TOWARD ABORTION

DEGREE FOR WHICH THESIS WAS PRESENTED DOCTOR OF PHILOSOPHY

YEAR THIS DEGREE GRANTED SPRING, 1988

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(SIGNED) E.M. ten Vergert.....

PERMANENT ADDRESS:

Kalkhofhorst 41.....

7531 EP Enschede.....

The Netherlands.....

DATED April 7.....1988

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled A SURVEY OF SCALING METHODS FOR THE MEASUREMENT OF ATTITUDES: A COMPARISON OF AMERICAN AND WEST-GERMAN ATTITUDES TOWARD ABORTION submitted by ELISABETH MARIA TEN VERGERT in partial fulfilment of the requirements for the degree of DOCTOR OF PHILOSOPHY in SOCIOLOGY.

Michael W. Gillespie
.....
Supervisor

Elisabeth Ten Vergert
.....

.....

J. Paul Johnston
.....

Clifford C. Clogg
.....
External Examiner

Date March 17, 1988

TO JOHANNES

ABSTRACT

A major theme of the present study has been issues in determining unidimensionality generally, and the comparability of attitude measures across populations, specifically. We addressed these issues by the application of both latent trait and latent class scaling models to the responses to seven abortion items in the 1982 NORC GSS and West German-ALLBUS combined files.

It was demonstrated that the non-parametric Mokken method provides a useful starting point in scale construction since it does not impose severe restrictions on the functional form of the item trace lines. The accompanying Mokken Test proved to be a useful means for developing equivalent attitude scales. We found that the seven items form a unidimensional scale in both countries and that four of these items constitute a scale that is robust across the two populations. A comparison of Americans and West Germans on this scale showed that Americans are somewhat more accepting of legalized abortion. We argued that the two-dimensional solution of earlier analyses is an artifact of the differences in the difficulty of the items.

The analyses of the parametric latent trait analyses (the one-parameter Rasch model, Birnbaum's two-parameter model, and McDonald's harmonic analysis) were not without problems. It was shown that these scaling models are not very practical to use in the case of a typical attitudinal data set like the abortion items.

Latent class scaling models were analyzed for the four robust items. The Proctor model augmented with an unscalable class provided the best fit for the American data; whereas the "pure" Goodman model was optimal for the West German sample. Results of the Mokken test and the Multi-group latent class models were compared.

Finally, log-multiplicative association methods were used to assign a metric to the categories of the ordinal robust Mokken scale. We concluded this study by showing how a cross-population LISREL model can be constructed that treats the four-item metric scale as the "reference indicator".

ACKNOWLEDGEMENTS

I would like to express my profound indebtness to my supervisor, Dr. M.W. Gillespie, for his many helpful comments as he read the manuscript at several stages, his continuous encouragement, and good humor over the past few years.

I am grateful to the other members of my committee, Dr. L.A. Hayduk, Dr. H. Krahn, and Dr. P. Johnston for their assistance and stimulating contributions to this thesis. I wish to thank, Dr. T. Taerum of the Statistical Consultant Service for the help in programming the scaling models. Margareth King who did an excellent editing job. Jos Ten Vergert who kindly donated her time and talent to assist in the research of some of the European literature.

Finally, I would like to thank my best friend Johannes Kingma. His support, optimism, concern, and encouragement during trying times were greatly appreciated.

Table of Contents

Chapter		Page
1.	GENERAL INTRODUCTION	1
	1.1 The Need For Non-Linear Models	3
	1.1.1 Introduction	3
	1.2 Classical True-Score Theory	4
	1.3 The Problem	9
	1.4 Justification of Research	15
2.	SOME BACKGROUND TO LATENT STRUCTURE ANALYSIS	18
	2.1 The Concept Of Attitude	18
	2.2 Classical View of Measurement	20
	2.3 Types of Scales	21
	2.4 Index Measurement	24
	2.4.1 Likert Scales	25
	2.4.2 Factor Analytic Scales	28
	2.5 Representational Measurement	29
	2.5.1 Thurstone Scales	31
	2.6 The Issue of Unidimensionality	33
3.	ASSUMPTIONS OF LATENT STRUCTURE-THEORY	36
	3.1 Introduction	36
	3.2 Latent Class Theory	37
	3.3 Latent Trait Theory	38
	3.3.1 Parametric Latent Trait Theory	38
	3.3.2 Nonparametric Latent Trait Theory	41
	3.4 Assumptions of Latent Structure Analysis	42
	3.4.1 Introduction	42
	3.4.1.1 Dimensionality	43
	3.4.1.2 Local Independence	43

3.4.1.3 Item Characteristic Curve 44

4. LATENT CLASS MODELS 45

4.1 Introduction 45

4.2 Deterministic Guttman Model 47

4.3 General Latent Class Model 48

4.4 Restricted Latent Class Model 50

4.4.1 Response Error Models 52

4.4.1.1 The Proctor Model 52

4.4.1.2 Item-Specific Error Rate Models ...55

4.4.1.3 Type-Specific Error Rate Model 56

4.4.1.4 False-Positive/False-Negative Error Models 57

4.4.1.5 Latent Distance Model 59

4.4.2 The Goodman Scale Model 61

4.4.3 The Dayton-Macready Model 62

4.5 Hierarchical Relations between the Latent Class Models 63

4.6 Multi-Group Latent Class Model 65

5. LATENT TRAIT MODELS 67

5.1 Introduction 67

5.2 The Non-Parametric Mokken Model 69

5.2.1 The Assumption of Double Monotony 70

5.2.2 Coefficients of Scalability 73

5.2.3 Test of Double Monotony 78

5.2.4 Cross-Population Comparisons of the Mokken Test 79

5.3 Parametric Models: Normal versus Logistic Ogive .81

5.3.1 Two-Parameter Logistic Model 83

5.3.2 Three-Parameter Logistic Model 86

5.3.3 One-Parameter Logistic Model 87

5.3.4 The Rasch versus Mokken Model 93

5.4 Latent Class Models and the Assumption of Double Monotony 96

5.5 Assigning a Metric to an Ordinal Scale 97

6. DATA AND METHODS 105

6.1 The Data 105

6.2 Method 108

6.3 Model Fit and Statistical Testing 110

6.4 Computer Programs Used for the Scale Analyses ..111

6.5 Estimation of Item and Person Parameters 112

6.5.1 Joint Maximum Likelihood Estimation 113

6.5.2 Conditional Maximum Likelihood Estimation 115

6.5.3 MOKKEN Program 117

6.5.4 PML 119

6.5.5 LOGIST 124

6.5.6 NOHARM 124

6.5.7 ANOASC 126

6.5.8 MLLSA 127

7. RESULTS OF THE "NON-PARAMETRIC ANALYSES" 128

7.1 Results of the Mokken Analyses 128

7.2 Results of the Latent Class Analyses 132

7.3 Results of the ANOASC Analyses 138

7.4 American and West-German attitudes toward abortion 143

7.5 Conclusion 151

8. RESULTS OF THE PARAMETRIC ANALYSES 152

8.1 Results of the Logist Analyses 170

8.2 Results of the NOHARM Analyses 174
8.3 Conclusion 177
9. SUMMARY AND DISCUSSION 179
9.1 Summary of Results 179
9.2 Discussion 183
9.3 Suggestions for Future Research 194
Bibliography 203
APPENDIX A 216
APPENDIX B 217
APPENDIX C 219

LIST OF TABLES

Table — Page

4.1 Conditional Probability of Positive Response for Four Items, Given Membership in One of the Five Latent Classes:Proctor's Stochastic Model. 53

4.2 Conditional Probability of Positive Response for Four Items, Given Membership in One of the Five Latent Classes: Item-Specific Error Rate Model. 55

4.3 Conditional Probability of Positive Response for Four Items, Given Membership in One of the Five Latent Classes: Type-Specific Error Rate Model. 57

4.4 Conditional Probability of Positive Response for Four Items, Given Membership in One the Five Latent Classes: False-Positive/False-Negative Error Rate Model. 59

4.5 Conditional Probability of Positive Response for Four Items, Given Membership in One of the Five Latent Classes: Latent Distance Model 60

4.6 Conditional Probability of Positive Response for Four Items, Given Membership in One of the Six Latent Classes: Goodman's version of Guttman's Deterministic Model 62

5.1 The Cross-Tabulation of Two Items. 73

7.1
The Distribution of the Abortion Items for Americans and West-Germans. 129

7.2
The H_i Values of the Abortion Items for Americans and West-Germans. 131

7.3
Impact of Dropping Items on the T-Statistic for the Mokken Test. 131

7.4
Chi-Square Values for Latent Class Models Applied to Four Robust Abortion Items for Americans and West-Germans. 134

7.5
Distribution of the Non-Metric and Metric Scale Scores for the Americans and West-Germans. 144

7.6
Estimated Scale Scores for the Abortion Attitude by Attitude on CHLDIDEL, CHLDNUM, PREMARSEX, XMARSEX, and HOMOSEX. 147

7.7
Results of the T and F-Test for the Instrumental Variables None, CHLDNUM, XMARSEX, and HOMOSEX. 149

8.1
Results from the Martin-Löf Test and Andersen Test of the Seven Abortion Items for the Americans. 154

8.2
Results from the Martin-Löf Test and Andersen Test of the Seven Abortion Items for the Germans. 155

8.3
Results from the Fischer and Schleibechner Test of the Seven Abortion Items for the American high/low score groups: 160

8.4
Results from the Fischer and Schleibechner Test of the Seven Abortion Items for the German high/low Score Groups. 160

8.5
Results from the Martin-Löf Test and Andersen Test of the Five Abortion Items for the Americans. 163

8.6
Results from the Martin-Löf Test and Andersen Test of the Four Abortion Items for the Germans. 165

8.7
Item Parameter Estimates of the Five Abortion Items for the Americans. 168

B.1
The P and P_o Matrix of the Seven Abortion Items for the Americans. 217

B.2
The P and P_o Matrix of the Seven Abortion Items for the West-Germans. 218

C.1
Latent Class Distribution Under the Models in Table 7.4 for the Americans. 219

C.2
Latent Class Distributions Under the Models in Table 7.4. for the West-Germans. 220

C.3
Distribution of Guttman Response Patterns According to Ordering (ANY, NOMORE, SINGLE, POOR, DEFECT, RAPE, HEALTH) for the Americans. 221

C.4
Distribution of Guttman Response Patterns According to Ordering (ANY, SINGLE, NOMORE, POOR, RAPE, DEFECT, HEALTH) for the West-Germans. 221

C.5
Results of the BINO test of the Seven Abortion Items for Americans in PML. 222

C.6
Results of the BINO Test for the Seven Abortion Items for the West-Germans in PML. 224

C.7
Results of the BINO Test for the Five Abortion Items for Americans in PML. 226

C.8
Results from the Fischer and Schleibechner Test of the Five Abortion Items for the American high/low Score Groups. 227

C.9
Results of the BINO Test for the Four Abortion Items for West-Germans in PML. 228

C.10
Results from the Fischer and Schleibechner Test of the Four Abortion Items for the West-Germans high/low Score Groups. 229

C.11
Item and Person Parameter Estimates of the Four Abortion Items for the West-Germans., Canada (Males). 229

C.12
Person Parameter, Item Parameter, and the Rasch Probability of a Positive Answer to the Items RAPE and POOR for the Americans. 230

C.13
LOGIST Item Parameters and Standard Errors of the Two-Parameter Model of the Seven Abortion Items for the Americans. 230

C.14
LOGIST and NOHARM Item Parameter Estimates of the Two-Parameter Model of the Seven Abortion Items for the Americans (N=702) and the West-Germans (N=1473). 231

C.15
NOHARM Item Parameter Estimates of the One Parameter Model of the Seven Abortion Items for the Americans. 232

LIST OF FIGURES

Figure — Page

1.1 Two Linear Model ICCs 7

5.1 ICCs for Four Items that Meet the Assumption of Double Monotony. 72

5.2 Normal and Logistic ICCs 82

5.3 Two-Parameter Logistic ICCs 85

5.4 Three-Parameter Logistic ICCs 88

5.5 One-Parameter Logistic ICCs 90

C8.1 Model with ABORT as reference variable 233

C8.2 Model with continuous variables and ABORT as reference variable 233

1. GENERAL INTRODUCTION

Two major objectives of science are description and explanation. As Torgerson states:

> "the principle of a science, other than the description of empirical phenomena, is to establish, through laws and theories, general principles by means of which the empirical phenomena can be explained, accounted for, and predicted" (Torgerson, 1958:1).

In sociology, for example, one could introduce a construct "attitude" and define it as a respondent's evaluation of a social situation. However, in order to make the construct more tangible, one has to translate it into a set of items; i.e., we have to operationalize the construct.

Science can be thought of as consisting of a theory (constructs) on the one hand and data (operationalizations of these constructs) on the other. Theory and data are linked by means of *correspondence rules* (Torgerson, 1958; Nagel, 1961). These rules of correspondence can also be considered as operational definitions of the constructs.

In the less advanced sciences, such as social sciences, constructs are often only loosely connected with their operationalizations. In these situations we have "a wealth of observables and certainly no lack of constructs. There is, however, a rather serious shortage of important connections" (Torgerson, 1958:5). In sociology, Schuessler (1982) demonstrates the prevailing chaos in operationalization and measurement of the construct "social life feelings". He found that about 1000 items appeared in over 100 "scales" of social life feelings (Schuessler,

1982:62).

Constructs are inventions and are established by observing that several indicators converge, i.e., covary. This implies that in delineating constructs, the researcher must already have some notion concerning the subject of her research. In fact, a researcher first decides on what kind of observations to take and then selects certain aspects of these observations for further analysis. The analyses themselves, however, are not dictated by the observations.

This point is emphasized by Coombs (1964), who makes a distinction between *observations* and *data*. Coombs argues that a researcher has data only when he has restructured the observations that will be analyzed. For instance, in the present study the answers to the items are coded "yes" or "no" and other responses such as "don't know" and "no answer" are ignored. The "yes" and "no" responses are interpreted as dominance data; e.g., a respondent is said to dominate an item which she answers positively (Coombs, 1964). The data dictate which measurement technique is to be used to represent the observations.

Usually, it is assumed that the construct is unidimensional; i.e., one construct dominates the data. This assumption can be tested by a series of unidimensional scaling models without changing the structure of the data. Two central questions can be raised then: first, are the data dominated by one latent trait or do they measure more than a single dimension; second, which scaling technique is

most suitable for the data?

1.1 The Need For Non-Linear Models

1.1.1 Introduction

Many of the recent developments in sociological methods focus on the use of multiple indicators (items) in the measurement of constructs or latent variables. This developments follows two traditions which we term, *Classical True-Score Theory* and *Latent Structure Theory*.

The critical basic assumption of these two traditions is that a set of items forming an instrument all measure just one thing in common. Further, it is assumed that a mathematical function describes the relationship between the observed variables (items) and the latent variable (construct). This relationship is a probabilistic one because both models assume that error is involved in the measurement process of the latent variable.

For each tradition, however, different mathematical models are used to describe the relationship between the observed and the latent variable, which are based on specific assumptions about the data.

In the next section, we shall introduce the classical true-score theory model, and in the Chapters 3, 4, and 5, we will examine the theory and logic of latent structure models. Based on the deficiences of the classical true-score theory model, we will argue that latent structure analysis

is a more appropiate technique to establish the unidimensionality of a set of *dichotomous* items.

1.2 Classical True-Score Theory

Models based on Classical True-Score Theory (CTST) are most frequently used in the measurement of unobserved or latent variables. In particular in attitude measurement and sociology, generally, these models have gained widespread acceptance.

CTST-models are based on weak assumptions, that is, the assumptions can be met easily by most data sets. Probably, because of the model's weak assumptions and because the model led to the formulation of a number of important statistics (Lord & Novick, 1968), the CTST-model became very popular in the social scienes.

The CTST-model is a very simple mathematical model that assumes a linear relationship between the observed and the latent variable (the true score). It is assumed that a person's score on an item is made up of two components: true score and error. The true score and the error are perceived as completely independent. The true score is viewed as unchanging from one set of items to a parallel set.

The error is considered to be unique to the specific measurement and to be entirely independent of the error that might equally be expected to appear on another measurement designed to assess that same latent variable of a person (Allen and Yen, 1979). The true score can never be directly

observed and is estimated by the observed variables. The observed score serves as an imperfect estimator of the "true score" because it is assumed to include error. As a consequence, CTST is a stochastic model. CTST is concerned with the reliability of measurements, and more specifically with the standard error of measurement.

Hambleton (1979; 1980) and Hambleton and Swaminathan (1985) outline the many shortcomings of CTST. One of its most severe limitations is that the item difficulty-the the proportion of respondents answering an item positively-serves as the basis for item analysis. Because item difficulty varies from group to group, it may be argued that:

> "Proportion of correct answers in a group of examinees is not really a measure of item difficulty. This proportion describes not only the test items but also the group tested" (Lord, 1980:35).

Another shortcoming of CTST is that comparisons of respondents on a latent variable measured by a set of items are limited to situations in which respondents are administered the same (or parallel) set of items. Thus, the number of positively answered items is not invariant from test to test and depends upon the particular set of items taken. In comparative research (the concern of the present study), item parameters that remain invariant across subpopulations and population parameters that remain invariant across a set of items are extremely important. Only when these conditions are met can the same set of items

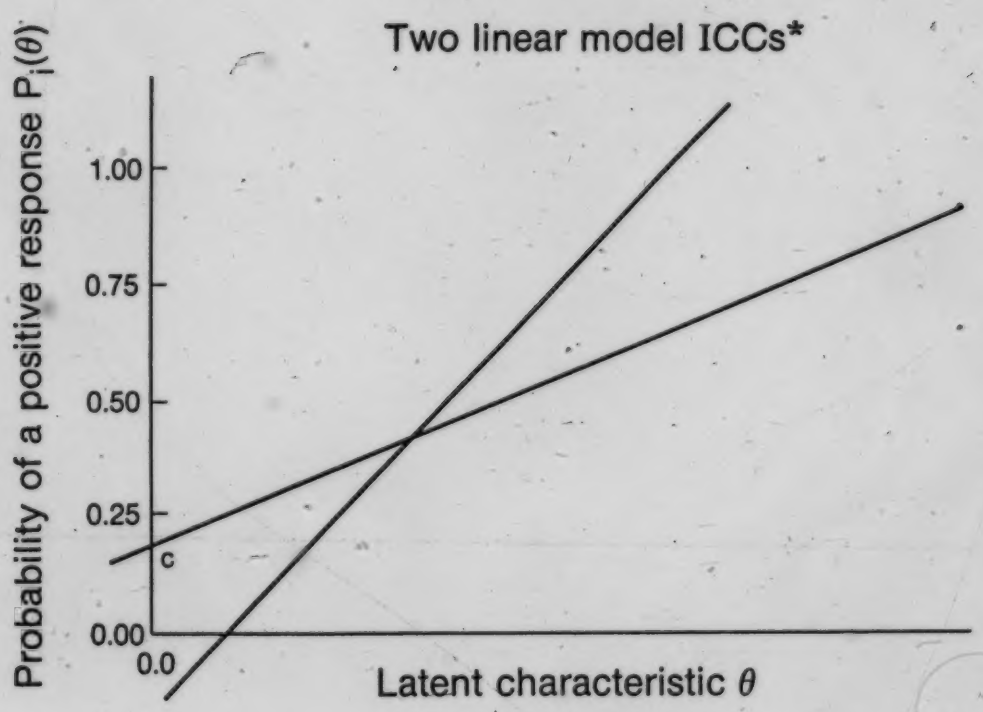
be used to compare subpopulations.

A further deficiency of CTST is that conventional item analysis does not describe how inferences from respondents' item responses can be made about a respondent's position on the latent trait.

Finally, variants of CTST are usually employed to evaluate the unidimensionality of a set of items, for instance, linear factor analysis and Likert's scale analysis. In the case of dichotomous item responses (the concern of the present study), this assumption of linearity may be violated because dichotomous items cannot be linearly related to continuous latent variables and this is particularly the case when items differ widely in difficulty levels (Hattie, 1985; McDonald, 1982; McDonald and Ahlawat, 1974; Hulin et al., 1983).

For instance, consider a dichotomous item and consider drawing a graph of the probability of getting a positive answer as a function of a latent variable such as attitudes toward abortion (see, Figure 1.1). We expect this probability to increase with the respondent's position on the latent attitude variable. However, as Figure 1.1 shows, since the measure is *discrete* and the latent trait is *continuous* we get probabilities less than 0 and greater than unity. As a consequence, common factor analysis will yield both *factors of content* and *factors of curvilinearity* whereas the latter factors are "mathematical artifacts" of the violation of the assumption of linearity (McDonald,

FIGURE 1.1



* An item characteristic curve (ICC) describes the relationship between the item and the latent variable

1985).

Classical test theory and associated procedures have failed to provide satisfactory solutions to the shortcomings mentioned above. Therefore, a good deal of interest among theoreticians and researchers has centered on latent class theory and latent trait theory in order to develop more appropiate theories of the measurement of latent variables.

Getting ahead of the chapters 3, 4, and 5, latent class-models are used for the analysis of discrete latent variables, whereas latent trait models can be used to examine continuous latent variables. Lazarsfeld refers to the two approaches as "Latent Structure Analysis" (Lazarsfeld, 1950). With respect to the four shortcoming of the CTST-model, latent structure analysis offers the following alternatives:

First, latent structure analysis has the attractive property of invariant item and person parameter estimation which is an important reason to use these models in comparative research instead of CTST.

Second, latent structure analyses provide methods of inferring a respondent's position on the latent variable from a set of items.

Third, latent structure analyses enable us to define mathematical functions that assume non-linear relationships between the observed and latent variables. Therefore, in the case of dichotomous variables (the concern of the present study), it is generally better to employ such models instead

of linear ones.

1.3 The Problem

There are four objectives to the present study: (1), to investigate (by using several latent structure models) whether the seven items on abortion, that have appeared in the 1982 NORC GSS and West German ALLBUS combined files, form one or two dimensions; (2), to see whether this set of items or a smaller subset comprise a scale that is robust across the populations under study; (3), to compare the Americans and West Germans on this robust scale; (4), to show how the robust scale can be used as the reference indicator (or indicators) in a LISREL model.

The abortion items constitute a particularly useful choice of an example for a study on unidimensionality since previous research has not resolved the issue of whether the different conditions under which abortion might be used constitute a unidimensional set of items. The conditions associated with the items are: (a) a strong chance of a serious defect in the baby (DEFECT),(b) the woman is married and does not want any more children (NOMORE), (c) the woman's own health is seriously endangered by the pregnancy (HEALTH), (d) the family has a very low income and cannot afford any more children (POOR), (e) the woman became pregnant as a result of rape (RAPE), (f) the woman is not married and does not want to marry the man (SINGLE), (g) the

woman wants it for any reason (ANY).[1]

Studies that suggest that the items form two scales (dimensions), rather than one, interpret these scales as: one made up of the three "easy" items, DEFECT, HEALTH and RAPE (abortion for "medical" or "hard" reasons), and the other scale made up of the NOMORE, POOR and SINGLE items (abortion for "social" or "soft" reasons).

The major empirical difference between the two sets of items is that the proportion of respondents who approve of abortion for medical reasons is substantially greater than the percentage who approve of abortion for social reasons. But whether this difference justifies the separation of the six items into two different measures is another question.

The question of whether the six abortion items are dominated by a single underlying dimension has been attacked directly and indirectly by researchers using a diverse array of methods. Some studies implicitly assume unidimensionality by simply summing the positive responses to each of the six items without presenting any evidence of scalability (Coombs and Welch, 1982; Finlay, 1985; Hall and Ferre, 1986). Other studies subject these items to a Guttman (1950) scale analysis (Arney and Trescher, 1976; Ebaugh and Haney, 1980; Granberg, 1978; Jones and Westhoff, 1973). All these studies report coefficients of reproducibility greater than .90 and conclude that the six abortion items are dominated by the same dimension. However, one can obtain very high values of

---

[1]The ANY item was introduced to the GSS in 1982.

the coefficients of reproducibility even though the items lack unidimensionality (Clogg and Sawyer, 1981; Guttman, 1950; Mokken, 1971; Nunally, 1978). High REP-coefficients are not compelling evidence of the unidimensionality of the abortion items.

Factor analysis of the six abortion items separates the medical and social items into two factors when one adopts the convention of extracting as many factors as there are eigenvalues greater than one (Arney and Trescher, 1976; Granberg, 1978). However, one could also interpret these results as an artifact of the differences in proportion of respondents who endorses the items. More specifically, the results of those factor analyses may be due to both factors of content and factors of curvilinearity.

Papers with a more methodological focus address the issue of the unidimensionality of the abortion items more directly. Clogg and Sawyer (1981) used a variety of latent class models and concluded that two dimensions, rather than one, underly these items. They base this conclusion on an eleven-class "biform" model. Ten of the classes correspondend to the scale types of a conventional Guttman scale analysis, and the eleventh represents a residual class of "intrinsically unscalable" individuals. Clogg and Sawyer argue that the ten scale types represent two orderings of the items. However, the difference between these two orderings has nothing in common with the previously discussed alternative to the undimensional interpretation:

abortion for medical reasons and abortion for social reasons. Moreover, the high percentage of cases in the unscalable class of Goodman's model presents a problem for comparative researchers, since removing these cases from the analysis obscures the nature of the populations being compared.

Mooijaart (1982) analyzed five of the six conditions by means of latent class analysis and concluded that a one-dimensional structure can be postulated to underly the items, i.e., a liberal versus a non-liberal attitude towards abortion.

Finally, Muthen (1982) analysis supports a two dimensional conclusion: the easy items measure attitudes toward the use of abortion for medical reasons, while the more difficult items measure attitudes toward the use of abortion for social rasons. Because Muthen's probit analysis posits a non-linear relationship between the item and the latent trait, it represents an improvement over conventional factor analysis. Nonetheless, probit factor analysis may fail to escape the consequences of the widely different p-values of the abortion items, because it may not capture completely the differences in the degree to which the item-trait relations depart from linearity. As Stinchombe (1982) has pointed out, one can draw on an infinitely large family of monotonic, non-linear relationships to model item-trait relationships. In line with these arguments, by specifying a particular form of a non-linear relationship

for *all* items, Muthen's probit analysis may also produce spurious factors.

In light of the above discussion, the Mokken method of scale analysis may provide a useful method for assessing the unidimensionality of the abortion items (Mokken, 1971; Kingma and Reuvekamp, 1986a, 1986b). As a latent trait method, the Mokken method represents a non-parametric stochastic extension of Guttman's method for scale analysis of items within a single population and, as such, places less stringent requirements on the data than Guttman's method and other parametric scaling methods. Instead, it requires only that the relationship be monotonic and that the tracelines, which describe the relationship between the item and latent variable, do not intersect. Therefore, the Mokken method may prove superior as a test for unidimensionality in the case of dichotomous items with widely different p-values.

In addition to this, neither the latent class models nor the other discussed models provide procedures either for reducing an item pool to a set of items that scales or for further reducing a set of scalable items to a subset that is robust across the populations studied.

This last feature of the Mokken Methods, the Mokken Test, is particularly important since it allows the researcher to test the extent to which the homogeneity of a set of items is uniform across the populations under study (Niemoller et al., 1980; Kingma and Reuvekamp, 1986b). This

last test may prove useful to the present study of establishing a scale of attitudes towards abortion that is robust across the American and West German populations.

Despite our comments regarding the more restrictive model specifications of the previously discussed parametric models, we want to emphasize that the Mokken method and test are regarded less as a competitor of these other methods than as a preliminary step in the construction of equivalent scales. We will show in the present study that if a particular set of items meets the Mokken criteria for a robust scale, more restrictive techniques such as the logistic test models (Birnbaum, 1968) can be used in order to see whether the items possess other psychometric properties that are common across the populations.

Finally, since the Mokken model assumes only an ordinal scale, the scale can be refined by assigning a metric to the categories of the cross-national scale. In the present study this will be accomplished by using the log-multiplicative methods discussed by Clogg (1982; 1984a, 1984b), Goodman (1984; 1987), Clogg and Goodman (1986) as well as by using logistic test models (Birnbaum, 1968) and McDonald's method of harmonic analysis (McDonald, 1967; Fraser, 1980).

Based on the results of these analyses we will decide which technique(s) is most useful in the establishment of metric cross-national scales for dichotomous data with widely different difficulty levels.

1.4 Justification of Research

Of course, the questions can be raised, why is it important to attempt to resolve the issue of unidimensionality? Why bother about robust scales?

First, many researchers maintain that the abortion items tap two attitudinal dimensions: attitudes toward the use of abortion for "medical" reasons and attitudes toward the use of abortion for "social" reasons. When these two subscales are produced by summing the hard and soft items separately, the researcher will likely find that the two scales will correlate somewhat differently with variables believed to be related to abortion attitudes. If the two subscales measure different dimensions, such differences raise substantive questions that warrant further empirical research. Why, for instance, would attitudes toward abortion for medical reasons correlate less with schooling than attitudes toward abortion for social reasons? On the other hand, if the two subscales measure a single dimension, differences in the correlations would represent an artifact of differences in the p-values of the items that comprise the scale. In this case, pursuing a substantive explanation of the difference in correlation would be a waste of time.

Second, further reducing a set of items that scales in individual populations to a core set of items that scales *across* the different populations may provide additional insight into the latent variable that the individual scale measure. For instance, which conditions of abortion comprise

a robust scale and why are other items not robust across the subpopulations?

Third, the development of an equivalent *scale* allows the researcher to separate cross-national differences in scores on the latent variable from cross-national differences in the item's location on the latent variable when either the non-parametric Mokken model or the one-parameter Rasch model is employed. Moreover, by evaluating different unidimensional scaling techniques that enable the establishment of robust scales, more insight will be gained into the usefulness of these techniques in attitude measurement in general.

Fourth, useful comparison by means of cross-population analysis LISREL models requires that the slope of the "reference indicator" [2] on the latent variable is invariant (Blalock, 1982:80-85).

Since the unobserved variable has no metric of its own, we may select one of the measured indicators to supply its metric. The selection of a reference indicator, however, poses a vexing problem for the use of LISREL in comparative research because valid comparisons require that the slopes (=unit of measurement) of the referent indicators are the same on the latent variable. This means that we need to

---

[2] A reference indicator is a variable whose "loading" or slope on the latent variable is set equal to one. The term "reference" is used because fixing the loading at one means that the metric of the latent variable will be the same as the metric of the reference variable. In addition, the loading of the other indicators which are estimated from the data also will be expressed in terms of the metric of the reference indicator.

develop a robust scale whose intervals between the scale scores are the same for both countries. Only then can "real" differences in the effects of several predictors of abortion attitude across research settings be studied.

2. SOME BACKGROUND TO LATENT STRUCTURE ANALYSIS

2.1 The Concept Of Attitude

In order to understand how attitudes can be measured, it is necessary to examine the concept of attitude. Like the majority of concepts in social sciences, the theoretical definition of the concept attitude is not well developed. Campbell (1950) argues that social scientists use eighty theoretical definitions of the concept of attitude. These eighty definitions, however, share the same operational definition. Since the present study is concerned about the *measurement* of attitudes, we will only discuss the integration of the theoretical and operational definitions of this concept.

According to Mead (1934), an attitude is an incipient act. It represents the initial stages of behavior. If we conceptualize attitude toward abortion in Median terms using the abortion *items* as our measure (or operationalization), we envision women having or not having abortions as the need arises, men supporting or not supporting that decision, political action or indifference toward moves to increase or decrease accessibility of legal abortion, etc.

Another reason for conceiving an attitude is the fact that social objects (attitudes) are more or less *general*. People respond both to these generalized object and to the specific contexts in which these objects appear. For this reason, the abortion items are indicators; i.e., the person

responds to both abortion, which is the focus of the question, and the context.

Green (1954) developed this argument further. He also argues that the concept of attitude does not refer to any specific act or response of a person, but that it is used to describe the consistency (or covariation) of a number of specific social situations (e.g., the abortion items) of the same general class (attitudes toward legal abortion). Therefore, it is assumed that when person A has a less favorable attitude toward legalization of abortion than person B, we mean that A's responses to a set of items about legalization of abortion are consistently less favourable than are B's comparable responses. Indeed, this transitivity holds only in the probabilistic sense and, more important, if the responses to the context of the items are equivalent.

From a comparative point of view, we emphasize the need to keep seperate the difference in the general attitude measured by a set of items from the difference in the proportion of respondents who endorses a specific items. While difference in the proportion of respondents who endorses a specific item represents an attitude change that is sociologically interesting, it nonetheless constitutes a different kind of difference that most researchers have in mind when they talk about differences in attitudes. The problem with using a single item to measure differences in general attitudes, then, is that responses to the item confound the general attitude with opinions about the

grounds that define the general context in which the item is couched.

In the traditional attitude scaling techniques (Thurstone, 1927; Likert, 1932; Guttman, 1944) as well as the more sophisticated Latent Trait Techniques (Rasch, 1961; Birnbaum, 1968; Lord, 1952; Clogg & Sayer, 1981) response consistency to a specific set of items about the latent variable (attitude) must be demonstrated.

2.2 Classical View of Measurement

The quantification of psychologial attributes like attitudes has been the subject of much debate. From the early twenties to 1940 Campbell's publications (Campbell, 1921; Campbell, 1928) were highly influential in what has been called *the classical view of measurement*. This classical view stems mainly from measurement in physics.

Campbell distinguished two kinds of measurement: *fundamental* and *derived* (Campbell, 1928:4). Fundamental and direct measurement are only possible when "addition" as a mathematical operation can be shown to be isomorphic with operations on the empirical objects. An example is the measurement of mass in which the "additive" operation consists of putting weights together in a balance. Only a few properties, such as weight and length, are measurable in this fundamental way. In the field of social sciences very few variables (if any) are measured by a fundamental procedure since most of these variables are unobservable and

consequently not physical or empirical quantities (Duncan, 1984a).

Derived measurements are based on fundamental measurement; i.e. they are derived by means of numerical laws relating fundamental measurements. For instance, density is measured by the ratio of mass to volume. Fundamental measurement and necessarily derived measurement result in ratio or interval scales. Both measurement procedures are rarely applied in social sciences (Torgerson, 1958).

2.3 Types of Scales

During the 1940's, social scientists felt that other approaches to measurement were needed which should not be limited to fundamental and derived measurement. Stevens (1951, 1959) broadened the definition of measurement considerably by including those empirical relations which can be quantified unambiguously. He interpreted measurement "as the assignment of numbers to quantities of the properties of objects in accordance with certain rules".

In establishing these rules for the use of a particular measurement the only procedure excluded is random assignment: "if there is no criterion for determining whether a given numeral should or should not be assigned, it is not measurement" (Stevens, 1959: 24). Steven's definition of measurement has become very popular and frequently appears in textbooks dealing with social measurement (see,

Torgerson, 1958; Bohrnstedt, 1983; Nunally, 1978).

Stevens introduced four types of scales, namely, nominal, ordinal, interval and ratio scales. In contrast to Campbell's restricted class of additive operations, he claimed the range of transformations under which a particular scale is invariant as relevant for measurement. The admissible transformations define the scale types and consequently there are as many scale types as classes of admissible tranformations.

These transformations, however, are limited by the data we are investigating; i.e., any transformation is permissable that preserves the empirical information contained in the scale. Thus, a scale may be transformed in any way that does not change any implications about the empirical system it represents.

Scaling then, can be defined as the assignment of numbers to properties of objects in such a way that the empirical relations between the objects are isomorphic with the corresponding relations between the numbers (Stokman and Verschuur, 1980). For example, in the case of the abortion items we assume a dominancy relation between the items. This means that when abortion item 1 is more often endorsed than abortion item 2, this must be represented by the number n_1 larger than the number n_2

We call our model a representation of the data, if the relationships among the items are properly reflected by the relationships among the numbers assigned to them based upon

a set of rules (measurement model). That is, we have measured the property (latent variable); we have established a scale of measurement.

Stevens introduced four types of scales:

Nominal scales represent the most unrestricted assignment of numbers to groups of objects (items). The numbers are used to represent categories in a classification and have no quantitative implications. The admissible transformations are only limited by the rule: "Do not assign the same numerals to different classes or different numerals to the same class" (Stevens, 1951:26). Nominal scales are of great importance in sociological research.[3]

Ordinal scales arise from the question of rank ordering and provide only information of the relative positions of items with respect to an attribute. Only the ordinal properties of the numbers are employed in measurement. The admissible transformations for these scales are the set of all order preserving transformations. Most of the scales used in the social sciences have ordinal properties exclusively.

Interval scales are based on an ordering of interval with respect to a certain property. The differences between the items (the intervals) with respect to the attribute are known, but no information is available about the absolute magnitude of the property of any subject. There is no knowledge of a "true" zero point. The scale remains

---

[3]See for criticism on Steven's definition of nominal scales (Duncan, 1984a; Torgerson, 1958).

invariant after the linear transformation y=ax+b.

Ratio scales are a particular type of interval scale with respect to a real zero point and the determination of successive equal intervals beginning at the zero point of the scale. Consequently, ratio scales allow only the similarity transformation y=ax. Ratio type scales are typically found in physics.

In the present study scaling models will be applied that assume ordinal and interval measurement levels of the latent variable, respectively.

In the next section we will show how scales can be evaluated by the use of two measurement techniques; representational and index measurement. In particular, index measurement is often used in attitude measurement. We will show that representational measurement is necessary to evaluate the consistency between the empirical observations of the property measured and the numerical scale values based on the scaling model.

2.4 Index Measurement

Index measurement, or measurement by fiat, is one of the weakest forms of measurement and is very commonly used in social sciences. Indices (scale values) can arbitrarily be defined by, for instance, the respondent's number of positive responses. However, this rule which leads to the assignment of scale values need not be consistent with the latent variable that is to be represented because no

empirical evaluations of the model are made as part of the measurement process (Yen, 1986; Dawes, 1972). In other words, index measurement techniques are somewhat ad hoc in obtaining scale values; "the prediction" that high total scores (e.g., total number of items endorsed) indicate more favorable attitudes towards abortion than lower scores must be true.

The level of measurement reached by an index measurement cannot be determined since the combination procedures to compute the indices do not refer explicitly or clearly to the variables that are to be presented, nor are they suggested by a specified relationship of the observations to such variables (Mokken, 1971). There is only a *one way* relationship between the empirical relations and the numerical relations, and, consequently, we cannot test (predict) if our numbers represent our empirical relational system (Andersen et al., 1983).

Index measurements are evaluated on the basis of their usefulness, i.e., their relationship with some external variable. Likert scales (Likert, 1932), reliability coefficients (Cronbach, 1951) and common factor analysis are examples of methods used to evaluate index measures.

2.4.1 Likert Scales

The Likert scale, also called "Summated rating model" (Likert, 1932) is a classic in the history of attitude measurement. The Likert scale was an alternative to the more

complicated Thurstone scale, being easier to construct and usually somewhat more reliable than the Thurstone scales. The Likert scale is mainly used for the scaling of individuals. Because the procedure is very simple we will discuss this model briefly and focus on the assumptions of the model.

A set of items is selected, each of which reflects desired behavior with respect to a particular attitude. Respondents are asked to indicate on a scale how much they agree or disagree with each statement. The person's attitude score is the sum of the scores on the separate items. High total scores indicate favorable attitudes toward the attitude variable and low scores represent unfavorable attitudes. Subsequently, the final scale can be subjected to a variety of tests by computing the scale reliability (Cronbach, 1951). It is assumed that the higher the reliability the more homogeneous the items are regarding the attitude variable.

However, the basic assumption of the Likert scale is rather strong. It is assumed that there is a linear relationship between each item and the underlying attitude variable; i.e., a more favorable attitude should produce a higher expected score on any particular item. As was demonstrated in section (1.2), this assumption is not realistic in the case of binary data.

In general, the Likert scale and the connected reliability coefficients are poor techniques to detect the

unidimensionality of a set of items. Particularly, these techniques are problematic when binary data are used. Hattie (1985), Guttman (1950) and Andrich (1985) argue that a high reliability coefficient is neither a sufficient nor a neccessary condition for unidimensionality. They conclude that the chief defect of these indices are their tendency to increase as the numbers of items increase. Conceptually, the unidimensionality of a set of items should be independent of its test length. As Cronbach (1951) puts it:

> "A gallon of homogenized milk is no more homogeneous than a quart" (Cronbach, 1951:325).

Another shortcoming of reliability coefficients is that they depend on the test as well as on the specific group of respondents. Samejima (1977a; 1977b) notes that it is easy to make a poor test look good "by using a heterogenous group of subjects and obtaining a large value of the 'reliability coefficient'. Similarly, we can make a good test look bad by using a homogeneous group of subjects" (Samejima, 1977b:196). Thus, reliability coefficients are dependent on a specific group of respondents and cannot be considered as parameters that are intrinsic to the test.

In addition to the reliability coefficients, factor analytic techniques are often used to find out whether the assumption of unidimensionality is violated or not. In the next section we will show that common factor analysis is not an appropiate technique to establish the unidimensionality of a set of items.

2.4.2 Factor Analytic Scales

Factor analytic techniques are often used in assessing the unidimensionality of a set of items. The purpose of factor analysis is to reduce a set of intercorrelated variables to a smaller set of unobserved factors or latent variables (Anderson et al., 1983). The researcher hopes that the analysis will reveal the attitude dimensions and how the items should be weighted to compose the scales. Often the percentage of variance explained by the first principal component is used as an index of unidimensionality.

However, despite the attractiveness and popularity of this technique, there are some general problems with factor analysis. Hattie (1985), for example, argues that many of the indices based on principal component analysis and exploratory factor analysis were shown to be inappropriate to answer the question about unidimensionality on a variety of grounds. These grounds include the statistical problems encountered in determining whether a correlation matrix has unit rank, whether to use component or factor analysis, determining the number of factors and the way of measuring communalities (Heise, 1974).

Furthermore, Stinchombe (1983), McDonald (1981; 1985). McDonald and Ahlawat (1974) and Hattie (1985) indicate problems may occur when the common factor method is used in the case of dichotomous items. These problems concern the assumption of a linear relationship between the dichotomous items and the latent variable. For a discussion of this

problem the reader is referred to section (1.2).

In sum, Likert scale analysis, factor analysis, and reliability coefficients are all problematic techniques in the assessment of the unidimensionality of a set of dichotomized items. Since they are all based on the assumptions of the classical true-score, they suffer from the same drawbacks as discussed in section (1.2)

2.5 Representational Measurement

To produce a scale whose level of measurement can be determined, it is necessary to employ representational measurement. In representational measurement, a *two-way* correspondence is established between empirical observations of the property being measured and relationships among the scale values produced by the measurement or scaling model (Yen, 1986; Dawes, 1972). Usually, several scaling models can be used to represent the empirical observations without changing the structure of the data.

An essential feature of representational measurement is the evaluation of the measurement model in terms of its internal consistency; that is, the match between the empirical observations and the model predictions based on the scale values. These scale values are based on the specific scaling model used to represent the observations.

The two-way correspondence necessary for representational measurement occurs if the observations determine the value on the measurement scale and if the

scale values may in turn be used to make predictions about the measured object. If the model is internally consistent, i.e., if the models fits the data, this fit is considered as evidence that a common latent variable may be measured by the observations. If the model does not fit the data we have to change our theory and respecify the assumed relationship between the latent variable and the items.

Only when representational measurement occurs is it possible to determine the scale type (i.e. level of measurement) of the measurement that has been produced by a particular measurement model. Therefore, representational measurement is a much stronger form of measurement than index measurement.

The concern for a measurement scale that is independent of the sample of persons measured was shared by investigators who developed scaling models according the techniques of representational measurement. For instance, both Thurstone's paired comparison scaling technique (Thurstone, 1927) and Guttman's deterministic scaling model (Guttman, 1950) consider the problem of invariance of parameter estimates. However, they did not represent these ideas mathematically in their models. The models applied in the present study are all based on the principles of representational measurement.

In the next section we will demonstrate how representational measurement works by discussing Thurstone's classic model of measurement.

2.5.1 Thurstone Scales

Thurstone's unidimensional scales for dichotomized data were originally developed to measure attitudes (Thurstone, 1927). His main concern was the search for a true metric, i.e., an attitude scale with equal intervals along a latent variable dimension. His basic model is produced by the technique of paired comparison based on the *Law of comparative judgement*.

According to this technique each person judges all possible pairs of statements and, for each pair, is asked which statement is more favorable toward the attitude object. The Law of comparative judgement was important because it provided a rationale for placing the statements on a psychological continuum (Edwards, 1953). This law assumes that the person's judgement is the outcome of a discriminal process. This process refers to different comparative judgements on successive occasions about the same pair of stimuli. Consequently, the discriminal process is not fixed due to random error.

The discriminal process can be expressed in a so-called discrimination distribution. It is assumed that these distributions for the statements follow a normal distribution and must overlap. The area of overlap represents a "confusion" area because it consists of two statements of which one item is sometimes called more favorable and sometimes the other. Because the discrimination processes are normally distributed, the

differences will also be normal. The mean of these differences is the best estimate of the scale value for the statements and can be estimated by calculating the proportion of times statement i was judged greater than statement j.

The Law of comparative judgement actually refers to series of judgements of a *single* observer. Therefore, simplifying assumptions are made. According to Thurstone:

> "it is probably safe to assume that the distribution of apparent weights for a group of subjects, each subject perceiving the weight only once, is also normal on the same scale" (Thurstone, 1927:52).

Only the values and positions of the statements are the focus of interest and not the individual differences between the subjects. After having obtained scale values for a set of statements, attitude scores for respondents can be measured.

Thurstone's effort to quantify attitudes by applying the method of paired comparison has been important in the historical development of the field, but Thurstone's scaling ideas are not used a great deal in current work.

First, the method of paired comparison is too cumbersome. Second, the assumptions of a non-monotonous normal distribution for the items will often be unrealistic in attitude measurement because each item receives agreement at only one zone of the attitude continuum (Nunally, 1978: 77-82).

Third, the assumption that the obtained scale values by the use of judges and respondents are the same may be too

strong because judges may think in a different way about the statements than the respondents whose attitudes are to be measured. However, by introducing this assumption, Thurstone realized already the significance of invariance of relative statement-values across persons with different attitudes (Andrich, 1978; Andrich, 1982; Duncan, 1984a). In this respect, he made the distinction between the construction and the application of a scale. If the scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it. This *necessity* of being able to estimate scale values independently of the characteristics of the calibration sample was later mathematically formalized in latent trait measurement models. In section 5.3.3. we will present this approach based on the psychometric work of the Danish mathematician Georg Rasch (1966).

2.6 The Issue of Unidimensionality

One of the most critical and basic assumptions of measurement theory is that a set of items forming an instrument measures just one thing in common (Lumsden, 1976; De Gruyter and Van der Kamp, 1984). Besides being the basis of most mathematical measurement models, there are several reasons why this assumption of unidimensionality is so crucial.

We agree with Lumsden (1976) that measurement begins with the conception of the measurable attribute. How can we

make any claims to measure if our measurement instrument has a number of different kinds of items based presumably on different attribute conceptions? Indeed the careful construction of a unidimensional set of items provides the basis for most scaling models.

Unidimensionality of a set of items is important when relating variables, ordering persons on an attribute, forming groups on the basis of some variable or making comments about differences between individuals or groups. As McNemar puts it:

> "Measurement implies that one characteristic at a time is being quantified. The scores on an attitude scale are most meaningful when it is known that only one continuum is involved. Only then can it be claimed that two individuals with the same score or rank can be quantitatively and, within limits, qualitatively similar in their attitude towards a given issue. As an example suppose a test of liberalism consists of two general sorts of items, one concerned with economic and the other with religious issues. Two individuals could thus arrive at the same numerical score by quite different routes. Now it may be true that economic and religious liberalism are correlated the meaning of scores based on such composite is questionable" (McNemar, 1946:368).

As previously mentioned, the measurement models discussed in the present study assume that a single latent variable underlies the items. Unidimensionality is defined then, as the existence of one latent variable underlying the data (Lord, 1981). Of course, this assumption cannot be met in any strict sense because there are always other latent variables that influence a person's response to a set of items. However, it is required for this assumption that a *dominant* latent variable underlies the items. In addition,

we note that the unidimensionality assumption must be checked carefully at each time point because changing social circumstances can affect the dimensionality of a set of items (Schuessler, 1982).

Recently, there has been an upsurge of interest in the topic of unidimensionality due to the increase of interest in latent structure models. As mentioned above, an advantage of these models is that they are (in principle) falsifiable. This means that it is possible to make predictions from the model and then check with observed data to see if these predictions are approximately right. In the Chapters 5 and 6, we will show that the assessment of the goodness-of-fit of these models is not without problems.

3. ASSUMPTIONS OF LATENT STRUCTURE-THEORY

3.1 Introduction

Latent structure analysis is based on two traditions; *latent class theory* and *latent trait theory*. Latent structure models were proposed to handle a number of problems which CTST has been unable to solve adequately

The formal roots of latent structure theory appear in a number of places. Lazarsfeld (1950), who conducted most of his research in the field of attitude measurement, was the first to introduce the term latent trait. Generally, a latent trait or unobserved variable can be identified by observed or manifest variables. In latent structure analysis a measurement model explicitly defines the relationship between the latent trait and the observable data.

The general notion behind these models is *the principle of local independence* (Lazarsfeld, 1950; Lazarsfeld and Henry, 1968). This basic concept of latent structure analysis means that any two items must be independent for given values of the latent variable. Lord and Novick (1968) give the definition of local independence more concrete meaning by saying that:

> "an individual's performance depends on a single underlying trait if, given his value on that trait, nothing further can be learned from him that contributes to the explanations of his performance. The proposition is that the latent trait is the only important factor and, once a person's value on the trait is determined, the behavior is random, in the sense of statistical independence" (Lord and Novick, 1968:538)

Although the principle of local independence is very explicitly treated in latent structure models, this does not imply that it is absent in the models associated to CTST. In applications of these latter models, a weaker alternative to the principle of local independence is considered.

The weaker form ignores moments beyond the second order and tests the unidimensionality of items by assessing whether the residual covariances are zero. Since it is possible for two items to be uncorrelated and yet not be entirely statistically independent (Hayduk, 1987:11), the strong principle is more stringent than the factor analytic principle that the residuals be uncorrelated.

3.2 Latent Class Theory

Initially, Lazarsfeld developed linear models for continuous latent variables. However, these models suffered from the same difficulties as the linear models discussed in the sections 2.4.1 and 2.4.2.; probabilities of a positive response less than zero or greater than one would occur. Therefore, Lazarsfeld substituted the linear model by a mathematically more tractable latent class model.

In the latent class model the distribution of the latent variable is assumed to be discrete rather than continuous. In other words, the latent class models assume nominal latent variables and are often referred to as qualitative models. It is supposed that within each of the latent classes the observed variables are mutually

statistically independent of one another. Interest in these models has increased with the development of estimation methods. Of particular interest for this study is the application of these methods to Guttman scaling (see Goodman, 1975; Proctor, 1970; Clogg, 1979; Clogg and Sawyer, 1981; Dayton and McReady, 1976; Clogg and Goodman, 1986).

3.3 Latent Trait Theory

3.3.1 Parametric Latent Trait Theory

The classic work, *A Theory of Test Scores*, (Lord, 1952) is generally regarded as the birth of latent trait theory. In it, Lord outlined a mathematical model defining the relationship between an observed item and the underlying trait, and developed methods for item and person parameter estimation. His model, termed the normal ogive model, assumes that the probabilities of responding positively to an item increase monotonically with increasing values on the latent trait. Lord claimed that his new theory

> "is more powerful ..... than direct application of the classical theory of errors starting with the broad assumption that test score and "true score" differ by normally distributed independent errors of measurement" (Lord, 1952:v).

Progress in latent trait theory in the 50's and 60's was painstaking slow, probably due to its mathematical complexity and a lack of convenient and efficient computer programs. Moreover, a lot of scepticism existed around this new technique.

In 1968, the appearance of Lord and Novick's (1968) *Statistical Theories of Mental Test Scores* (with four chapters by A. Birnbaum) had a major impact on the development of latent trait theory. The book discusses the statistical theory of the logistic model extensively. Logistic model proved to be much simpler mathematically than the normal ogive model. This relative simplicity of the logistic model arises from the fact that the distribution of responses belongs to the exponentional family.

In the next section, we will discuss three different logistic models. If these models fit the data, equal interval scales according to the requirement of representational measurement have been established. Latent trait analysis has been strongly advocated for the analysis of cognitive and ability tests. The application of latent trait theory to attitude measurement lags far behind, however, although such application has been recommended in the methodological literature (Duncan 1984b, 1985; Reiser, 1981).

As previously mentioned, latent structure models are based on the strong notion of "local stochastic independence". However, models based on CTST are based on the weaker form of this principle, i.e., the partial correlations of a set of items are all zero if the common factors are partialled out. McDonald (1981a, 1982, 1985) claims that models based on the weak form of the assumption better address the issue of dimensionality than models based

on the strong assumption. In that respect, he describes two general approaches to the estimation of the parameters of latent trait models.

The first, the fixed trait approach, is based on the strong assumption of local independence and estimates jointly the *n* item parameters and the *N* ability parameters of the respondents by means of a maximum likelihood procedure. However, most computer programs written to implement this group of methods do not give suitable measures of the misfit of the model to the data (McDonald, 1985:210). (This problem will be discussed further in Chapters 5 and 6.)

The second, the random traits approach, treats the ability of the examinees as a random variable and does not try to estimate its values in the sample. It is assumed that the latent trait has a normal distribution. McDonald (1967) showed that we may approximate the normal ogive model (as previously mentioned this model is almost the same as the logistic model) by means of a polynomial series, using harmonic analysis. Fraser's program NOHARM- (Normal Ogive Harmonic Analysis Robust Methods)- (Fraser, 1980), was developed to fit this model. McDonald claims that a major advantage is the fact that this model yields residual covariances of the items that can be considered as measures of the departure in the sample of the assumption of local independence, in its weak, correlational sense. If the residuals are too large the assumption of unidimensionality

is rejected.

3.3.2 Nonparametric Latent Trait Theory

The Mokken method (Mokken, 1971; Kingma & Reuvekamp, 1986a) represents a nonparametric stochastic extension of Guttman's method for the scale analysis of items within a single population (Guttman, 1950). It is non-parametric in the sense that it imposes no fixed form (for instance logistic) on the function of the relationship that links the item to the latent variable. Instead, it requires only that the relationship be monotonic and that the trace lines which describe this relationship do not intersect. In addition, the Mokken method provides procedures for reducing an item pool to a set that scales

Developed some ten years later than the Mokken method, the Mokken test provides a statistical basis for deciding whether a set of items that is scalable according to Mokken's criteria, is robust across different populations (Niemöller, et al., 1980; Kingma and Reuvekamp, 1986b). From a comparative point of view (the concern of the present study), this last test is extremely useful since it will allow us to determine whether the pattern of relationships among items holds across two or more populations.

In light of the above discussion it may be argued that the Mokken model assumes only an ordinal level of the latent variables; the distance that separates the scale score of the items plays no rule in the outcome of the Mokken test.

However, the ordinal scale can be refined by assigning a metric to the ordinal scale values. This can be accomplished by using either the log-multiplicative methods discussed by Clogg (1982; 1984a; 1984b), Goodman (1984; 1987) and Smith and Garnier (1987) or the models of Rasch (Rasch, 1960; Duncan, 1984b).

Substitution of metric values for the ordinal scale values would allow us to measure the differences between the scale scores of the items.

3.4 Assumptions of Latent Structure Analysis

3.4.1 Introduction

Latent structure models and CTST-models differ in the assumptions made. Latent structure models are based on stronger assumptions about the data, which will lead to stronger results if the model fits the data. By specifying the assumptions one is willing to make about the data different latent structure models can be formed. The adequacy of the assumptions can be checked for any data set assuming that N is sufficiently large.

In general, three important assumptions arise in connection with latent structure analysis: dimensionality of the latent space, local independence, and item characteristic curve (Lord & Novick, 1968; Lord, 1980; DeGruyter and VanDerKamp, 1984);

3.4.1.1 Dimensionality

Latent structure analysis assumes that one latent trait defines the latent space i.e. a single latent trait "explains" or "accounts" for a person's response probability to a set of items.

Multidimensional trait models do exist but they have not yet been developed so as to be applicable on a large scale for practical purposes. Multidimensional models have been discussed by Samejima (1974).

Despite the importance of the assumption of unidimensionality in latent structure analysis, there is "not an accepted and effective index of the unidimensionality of a set of items" (Hattie, 1981:2). For a review of the various methods for determining the unidimensionality of a set of items, we refer to Hattie (1984; 1985).

3.4.1.2 Local Independence

As discussed in Sections 3.1. and 3.1.1, the principle of local independence is an essential assumption in latent structure analysis. This principle states that item responses are conditionally independent given the same value on the latent variable, so that the probability of joint response for persons with the same value equals the product of their marginal probabilities of response. In the case of *one* latent variable, the assumption of local independence corresponds to the assumption of an undimensional space (Lord, 1980).

However, the converse is not true; the principle also holds for zero or more than one latent trait.

3.4.1.3 Item Characteristic Curve

An item characteristic curve is a mathematical function that relates the probability of a positive response on an item to the latent variable measured by the item set that contains it. To put it differently, it is the nonlinear regression function of the item score on the latent trait measured. The probability of a respondent providing a positive answer to an item depends only on the form of the item characteristic curve; therefore, "it is independent of the distribution of examinee ability in the population of examinees of interest" (Hambleton et al., 1978). This invariance property is one of the useful features of latent structure analysis. As previously discussed, most parametric latent trait models belong to the exponentional family. This means that the item characteristic curve of these models share the same logistic form.

4. LATENT CLASS MODELS

4.1 Introduction

Latent structure analysis has been developed following two different approaches; one line which has resulted in the development of latent trait models (Lord, 1952, 1953, 1980; Rasch, 1966) and a second line of investigation which has stressed the application of latent structure concepts with qualitative data. This latter kind of technique resulted in the development of "latent class models" (see e.g., Lazarsfeld, 1950; Lazarsfeld and Henry, 1968; Goodman, 1974a, 1974b). Models of this kind differ from the latent trait models in that they treat the latent variable as a set of discrete classes rather than as a continuous variable. On the other hand both approaches are unified by the fact that both take local independence as the basis for measuring the latent variable.

Latent class models offer a way to test a wide range of hypotheses attempting to explain the relationship among a set of items in terms of categorical latent variables. For instance, latent class models are used in stratification research (Clogg, 1981), educational research (Haertel, 1984) and scale analyses (Clogg and Sawyer, 1981).

Latent class models can be divided into "general unrestricted latent class models" and "restricted latent class models". In the former, variations can be produced by altering the number of latent classes whereas in the latter,

restrictions can be placed on both the latent classes and the conditional probabilities of the general model. The hypothesis of interest determines the restrictions that are imposed on the parameters of the latent class model. If the model fits the data (i.e. the assumption of local independence is not violated), the probability of a given response on any one item is independent of the probability of any given response on any other items *within* any given latent class. To put it differently, the relationships among the items can be explained by the latent classes of the latent variable.

In the following, we will consider latent class scaling models that are explicitly developed as stochastic extensions of Guttman's deterministic scaling model (Guttman, 1950). For this reason, we begin with a brief overview of Guttman's model. Next, we consider the General Unrestricted Latent Class Model and the Restricted Latent-Class Model. After that we present the scaling models that are used in the present study and which all belong to the category of Restricted Latent Class Models. Finally, the technique of Simultaneous Latent Class Analysis will be outlined. In these models the same criterion of unidimensionality is used which Guttman (1950) adopted in his classic work on the scaling of dichotomous items; i.e., endorsement of an item of a given level of difficulty implies the endorsement of less difficult items.

4.2 Deterministic Guttman Model

To illustrate the use of the models and methods to be discussed in the next section we first consider the situation where there are four dichotomous variables, say, A, B, C, and D, and let (i, j, k, l) denote the response pattern where A is at level i, B is at level j, C is at level k, and D is at level l (for i= 1,0; j=1,0; k=1,0; l=1,0). Now assume that we use the item marginals (p-values) to order the items. Under Guttman's deterministic model (Guttman, 1950), only the following set of response patterns is allowed:

$$T = \{(1, 1, 1, 1), (1, 1, 1, 0), (1, 1, 0, 0), (1, 0, 0, 0), (0, 0, 0, 0)\}. \quad (1)$$

Note that although the number of possible response patterns (i, j, k, l) pertaining to the joint variable (A, B, C, D) equals 2^4, the number of consistent response patterns for k Guttman scalable items equals k+1. These response patterns in T will be called *scale-types* and the numbers t=1,.....,5 correspond to these scale types. For instance, $t_1=(1, 1, 1, 1)$, $t_2=(1, 1, 1, 0)$, $t_3=(1, 1, 0, 0)$, $t_4=(1, 0, 0, 0)$, $t_5=(0, 0, 0, 0)$.

Guttman's scale model assumes *strict* ordinality of the items in terms of the marginal distributions of yes/no responses to the individual items. The respondents can be ordered in terms of the number of positive responses. It is

assumed that no response pattern will be observed that is inconsistent with one of the k+1 scale types. In empirical research, however, measurement error is involved and response patterns other than Guttman's scale types will usually occur.

To circumvent this problem, various *probabilistic* extensions of Guttman scaling have been developed which allow for inconsistent, or non-scale type responses by incorporating the concept of "response error". These models are considered below.

4.3 General Latent Class Model

Goodman (1974a) specified a general, unrestricted latent structure model developed by Lazarsfeld (see Lazarsfeld and Henry, 1968) that is based on only one assumption, i.e., the principle of local independence. Goodman's model is developed for use in the analysis[4] of contingency tables involving the cross classification of more than one dichotomous or polytomous variables. The researcher tries to find a latent variable(s) that can "explain" the relationship between the items.

Consider again the case of the four manifest variables A, B, C, and D, with observed frequencies in the ABCD table. Let X denote the unobserved or latent variable in a particular model, and let π^X_t refer to the probability that a respondent will be a member of the t-th class of X.[4]

---

[4]In the present study we will adopt the notation used by Goodman (1974a; 1974b) who developed maximum likelihood

Next let $\pi^{\bar{a}\,x}_{i\,t}$ denote the *conditional* probability that an individual is in class i of variable A, given that she is in class t of variable X, and let the conditional probabilities $\pi^{\bar{b}\,x}_{j\,t}$, $\pi^{\bar{c}\,x}_{k\,t}$, and $\pi^{\bar{d}\,x}_{l\,t}$ be defined similarly (for i=1,0; j=1,0; k=1,0; l=1,0). (When superscripts have both barred and unbarred symbols, the unbarred symbols denote variables on whose levels we condition).

Further we suppose that the latent variable X with T classes accounts for the original ABCD association so that the following relationships hold:

$$\pi_{ijkl}=\Sigma\pi A_i B_j C_k D_l X_t, \qquad (2)$$

where

$$\pi A_i B_j C_k D_l X_t = \pi^{x}_{t}\,\pi^{\bar{a}\,x}_{i\,t}\,\pi^{\bar{b}\,x}_{j\,t}\,\pi^{\bar{c}\,x}_{k\,t}\,\pi^{\bar{d}\,x}_{l\,t} \qquad (3)$$

Equation (2) expresses the assumption that the probability of response (i,j,k,l) is obtained by collapsing over the categories of the latent variable X which is only indirectly observed in the ABCD table. Expression (3) denotes the probability that a respondent will be in cell (i,j,k,l) of

---

'(cont'd) estimation procedures for the parameters in latent class structures, and Clogg (1977) who has implemented these procedures in the MLLSA computer program. The presentation of the model restrictions is the way in which they are implemented in MLLSA.

the *unobservable* cross-classification of the response items and X (A,B,C,D,X). Equation (3) states that the respondent can be classified into T mutually exclusive and exhaustive latent classes; each respondent is in one and only one latent class, and all respondents are in some class of X. For instance, $\pi^{\bar{a}\,x}_{i\,t}$ is the conditional probability that a respondent is at level i of variable A, given that she is in class t of variable X.

Note that equation (3) is based on the "principle of local independence"; i.e., the variables A, B, C, and D are conditionally independent *within* the t-th latent class of X. Stated differently, the latent variable X (with possibly many categories) explains the relationships between the observed variables A, B, C, and D. Equations (2) and (3) define the general latent class model (see Lazarsfeld and Henry, 1968; Clogg, 1981; Clogg, 1977).

4.4 Restricted Latent Class Model

As noted above, variations in latent class models can be produced by both altering the number of latent classes and by imposing restrictions on the probabilities of the general model. These restrictions can constrain the probabilities associated with latent classes but also the conditional response probabilities associated with latent classes. The models considered in the present study differ in two respects: (a) the restrictions placed on the conditional probabilities of response, given membership in a

particular latent class; and (b) whether or not the model specifies a class of inherently unscalable individuals.

One feature of the use of latent class models with dichotomous variables is that a restriction imposed on one conditional probability automatically imposes a restriction on the other because conditional probabilities have to sum to unity. The restrictions that can presently be considered with MLLSA are of the following types:

1. Equality restrictions on the conditional probabilities; for instance, $\pi^{\bar{a}}{}_{1}{}^{x}{}_{1} = \pi^{\bar{b}}{}_{1}{}^{x}{}_{1}$. Since the conditional probabilities sum to unity, this restriction also implies the following restrictions: $\pi^{\bar{a}}{}_{0}{}^{x}{}_{1}=1-\pi^{\bar{a}}{}_{1}{}^{x}{}_{0}$, $\pi^{\bar{b}}{}_{0}{}^{x}{}_{1}=1-\pi^{\bar{b}}{}_{1}{}^{x}{}_{1}$, we would also have $\pi^{\bar{a}}{}_{0}{}^{x}{}_{1}=\pi^{\bar{b}}{}_{0}{}^{x}{}_{1}$ and so on.
2. Restriction of parameter to fixed constants; for instance, $\pi^{\bar{a}}{}_{1}{}^{x}{}_{1}=1$. This restriction implicitly restricts $\pi^{\bar{a}}{}_{0}{}^{x}{}_{1}=0$.
3. Equality restrictions on the latent class proportions: $\pi_1=\pi_0$.
4. Restrictions that form a combination of type 1-3 (See Clogg, 1977).

In the following, we will describe how various model specifications can be translated into constraints on the conditional probabilities. These constraints must be used in the computer program MLLSA for latent classes (Clogg, 1977), in order to get parameter estimates for the different models.

4.4.1 Response Error Models

The first type of model that allows for measurement error assumes that the entire population is "intrinsically" Guttman scalable and that the conditional probabilities are restricted to reflect assumptions about response error rates for respondents in a particular class. In the following, we will discuss those response error models that are applied in the present study. These models are ordered in the sequence of imposing fewer restrictions on the response error rates.

4.4.1.1 The Proctor Model

The strictest and simplest form of the type discussed in this section is that of Proctor (1970). His model assumes that only *one* "error" parameter is needed to describe the conditional probability of a positive response to an item and this parameter is independent of the items as well as of the latent classes.

Proctor's model assumes further that for an analysis of k items, k + 1 "true" types of individuals can be distinguished that correspond to the pure scale types under Guttman's deterministic model. An individual in the t-th latent class (or t-th scale type) is assumed to respond to each item with probability a of error, and this response error probability is assumed to be statistically independent of the probability of response error to any of the other items. An individual belonging to the t-th scale type can therefore make j response errors (for $0 \leq j \leq k$) with probability $(1-a)^{k-j}a^{j}$.

Table 4.1 presents the equality restrictions used in the five-class latent structures to define Proctor's model. Let rows one through four represent items 1 through 4, where item 1 is least difficult and item 4 is most difficult. Let columns one through five present the scale types I (1,1,1,1) through V (0,0,0,0), where scale type I represents the "highest" score on the latent variable and scale type V the "lowest". Let *a* represent a conditional probability that can range from between 0 and 0.5. The conditional probability of response for the Proctor models is presented in Table 4.1.

Table 4.1: Conditional Probability of Positive Response for Four Items, Given Membership in One of the Five Latent Classes: Proctor's Stochastic Model.

	Scale Type				
Item	highest I	II	III	IV	lowest V
1 easiest	1-a*	1-a	1-a	1-a	a
2	1-a	1-a	1-a	a	a
3	1-a	1-a	a	a	a
4 hardest	1-a	a	a	a	a

*a represents a quantity such that $0 \leq a < .5$

Because item responses are assumed to be conditionally independent given latent class, the conditional probability of any pattern of correct or incorrect responses, given that the respondent is in the

t-th latent class, is just the product of the conditional probabilities of their separate occurences. Thus, as is shown in Table 4.1, for a 4-item scale an individual that is intrinsically a member of scale type t_1, with expected response pattern (1,1,1,1), would have a probability of:

$(1-a)^4$ of giving response pattern (1,1,1,1);
$(1-a)^3a^1$ of giving response pattern (1,1,1,0), (1,1,0,1), (1,0,1,1) or (0,1,1,1);
$(1-a)^2a^2$ of giving response pattern (1,1,0,0), (1,0,1,0), (1,0,0,1) or (0,1,1,0);
$(1-a)^1a^3$ of giving response pattern(1,0,0,0), (0,1,0,0), (0,0,1,0) or (0,0,0,1);
a^4 of giving response pattern (0,0,0,0).

The exponent of a in each term refers to the number of modifications in the scale type response pattern that are required to produce the observed pattern. When a is high it is doubtful whether every respondent in a population in fact belongs to one of the k+1 scale types. Consequently, the scalability (unidimensionality) of the items can be questioned. When a=0 (no response error), Proctor's model equals Guttman's model.

Proctor's model is severely restricted, and thus unlikely to fit most data sets very well. Clogg and Sawyer (1981), Dayton and Macready (1980), and Rindskopf (1983) discuss several models that relax the assumption that response errors rates be identical for each item

and each latent class.

4.4.1.2 Item-Specific Error Rate Models

Item specific-error rate models allow the conditional probabilities to vary from item to item within a scale type. Thus, the error rate a_i is dependent upon the item but does not vary across the scale types of individuals (the latent classes). Again, it assumed that the conditional probability of response error to one item is statistically independent of the probability of response error to any of the other items for members of the same scale type. The restrictions that define this type of model are presented in Table 4.2.

Table 4.2: Conditional Probability of Positive Response for Four Items, Given Membership in One of the Five Latent Classes: Item-Specific Error Rate Model

	Scale Type				
Item	highest I	II	III	IV	lowest V
1 easiest	1-a*	1-a	1-a	1-a	a
2	1-b	1-b	1-b	b	b
3	1-c	1-c	c	c	c
4 hardest	1-d	d	d	d	d

*a, b, c and d represents a quantity such that $0 \leq a, b, c, d < .5$

For example, consider an individual in scale type II, with expected response pattern (1,1,1,0). According

to the principle of local independence, we can calculate the probability that an individual would make the response pattern (0,1,0,0) given that she is in scale type II by simply multiplying the item marginals:

$$a(1-b)c(1-d)$$

where $a \neq b \neq c \neq d$.

Thus, it is expected that the items 1 and 3 will be answered incorrectly and that the items 2 and 4 will be observed correctly. Note that under this model k error rates are to be estimated (one for each item) and that in Proctor's model these k parameters are assumed to have the same value. If the item specific error-rate model fits the data, the error rates vary for each item and the model parameters tell us to what extent they vary within each scale type.

4.4.1.3 Type-Specific Error Rate Model

Another response error model allows the conditional probabilities to vary for each scale type and not by item. This type-specific error rate model is presented in Table 4.3.

Table 4.3: Conditional Probability of Positive Response for Four Items, Given Membership in One of the Five Latent Classes: Type-Specific Error Rate Model

	Scale Type				
Item	highest I	II	III	IV	lowest V
1 easiest	1-a*	1-b	1-c	1-d	e
2	1-a	1-b	1-c	d	e
3	1-a	1-b	c	d	e
4 hardest	1-a	b	c	d	e

*a, b, d and e represent quantities such that $0 \leq a, b, c, d, e < .5$

For instance, assume that a respondent is a member of scale type II (1,1,1,0). The probability of observing response pattern (0,1,0,0) under this models is:

$$b(1-b)(b)(1-b)$$

The main problem of the type-specific error rate model is that the error rates are the same for both false-positive and false-negative answers. The model discussed in the next section will circumvent this problem.

4.4.1.4 False-Positive/False-Negative Error Models

Dayton and Macready (1976, 1980) consider models that assume *two* different types of errors. The

false-positive type of error *b* occurs if an observed response is positive but the relevant scale type calls for a negative answer. A false-negative type of error *a* occurs when a negative response has taken place but the scale type calls for a positive answer. Either the first type or the second type of error may occur, for instance, when a respondent is careless or distracted. In the context of ability testing the second type of error refers to respondents that answer an item positively due to guessing or cheating. In this case, the interpretation of the false-positive type of error is similar to the interpretation of the *c* parameter in a three parameter logistic model (See section 5.3.2). A discussion of the comparability of the error parameters in the Dayton and Macready models and the *c* parameter in the three-parameter logistic model is provided by Van der Linden (1978). The conditional probability of response for the false-positive /false-negative error model are presented in Table 4.4.

Table 4.4: Conditional Probability of Positive Response for Four Items, Given Membership in One of the Five Latent Classes: False-Positive/False -Negative Error Rate Model

	Scale Type				
Item	highest I	II	III	IV	lowest V
1 easiest	1-a*	1-a	1-a	1-a	b
2	1-a	1-a	1-a	b	b
3	1-a	1-a	b	b	b
4 hardest	1-a	b	b	b	b

*a, and b represent quantities such that $0 \leq a, b < .5$

It is assumed that the probability of a false-positive error b is constant across both items and scale types (excluding, of course, scale type I, where a false-positive error is by definition impossible); similarly for a false-negative error a (which is impossible in scale type V).

For instance, a respondent who is a member of scale type II with expected response pattern (1,1,1,0), would have an observed response pattern (0,1,0,0) with a probability:

$$a(1-a)(a)(1-b)$$

4.4.1.5 Latent Distance Model

Finally, the latent distance model (Lazarsfeld and Henry, 1968; Chapter 5) allows error rates to vary from item to item, and also allows the false-positive error

rates to differ from the false-negative error rates. (Note that in the former model the two error types were not allowed to differ across the items).

In order for the model to be identified, Lazarsfeld and Henry (1968) assume that the error rates of false-positive and false-negative responses are equal for the most "extreme" items (i.e., the easiest and the most difficult item). These two items serve as "anchor" for the scale. Each intermediate item, then, has two error rates: one for the false-positive responses and one for the false-negative responses. To describe the latent distance model in terms of conditional probabilities. We define the following restrictions depicted in Table 4.5 in describing the conditional probabilities of the latent distance model.

Table 4.5: Conditional Probability of Positive Response for Four Items, Given Membership in One of the Five Latent Classes: Latent Distance Model

	Scale Type				
Item	highest I	II	III	IV	lowest V
1 easiest	1-a*	1-a	1-a	1-a	a
2	1-b	1-b	1-b	c	c
3	1-d	1-d	e	e	e
4 hardest	1-f	f	f	f	f

*a, b, c, d, e and f represent quantities such that $0 \leq a, b, c, d, e, f < .5$

4.4.2 The Goodman Scale Model

Goodman (1975, 1979a) developed a latent class model that divides the respondents into two groups: a scalable- and an intrinsically unscalable class. It is assumed that responses of the scalable class *always* conform to the scale types expected under Guttman's model and members of the unscalable class are those who respond to the items either randomly or who order the items differently than do the majority of the respondents.

Note that Goodman's model assumes that there are k+2 latent classes. The scale types constitute k+1 latent classes and the members of these classes are assumed to respond on each item with a probability of one or zero. The final class represents the random component; the members of this class are deemed "inherently unscalable".[8] The restrictions that define Goodman's model are presented in Table 4.6.

---

[8] Goodman (1975, 1979a) used the "quasi-independence" concept when he introduced his scaling model. This term refers to the fact that only a subset of the 2^k response patterns (the non-scale type patterns) are assumed to be mutually independent. Usually, quasi-independence is used as a baseline model in stratification research where the class of "movers" is assumed to be statistically independent of the originating class.

Table 4.6: Conditional Probability of Positive Response for Four Items, Given Membership in One of the Six Latent Classes: Goodman's Version of Guttman's Deterministic Model.

		Scale Type				
		highest				lowest
Item	0	I	II	III	IV	V
1 easiest	a	1.0	1.0	1.0	1.0	0
2	b	1.0	1.0	1.0	0	0
3	c	1.0	1.0	0	0	0
4 hardest	d	1.0	0	0	0	0

*a, b, c, d, represent quantities such that $0 \leq a, b, c, d < 1.0$

Table 4.6 shows that the restrictions on the conditional probabilities of the latent classes I to V are the same as those associated with the Guttman model. Note also that the conditional probabilities associated with the latent unscalable class (latent class 0) are unrestricted because the responses of the members of this class to the k items are viewed as mutually independent.

4.4.3 The Dayton-Macready Model

Dayton and Macready (1980) developed a model that combines the features of the Goodman's scale model and the item response error models. The model assumes that a latent unscalable class exists (as in Goodman's model), but relaxes the assumption that members of the scalable class respond error-free. To describe the error structure of the scalable class in this model, we can use the various response models

described above.

Dayton and Macready (1980) also show how a test of significance can be constructed to assist in deciding on the necessity for including an intrinsically unscalable class in the model and alternatives including error responses for scale types.

4.5 Hierarchical Relations between the Latent Class Models

Nested or hierarchical models are straightforward; i.e., nesting simply implies that one of the tested models is a subset of the other. These subsuming relations among models are referred to as nested or hierarchical models (Fienberg, 1983; Rindskopf, 1983). In nested or hierarchical models, the chi-square of the more restrictive model (X^2_i) is compared to the chi-square of the less restrictive model (X^2_j); $X^2_i - X^2_j$ and has a chi-square distribution with degrees of freedom equal to $df_i - df_j$.

For instance, the item-specific error rate model is less restrictive than the Proctor model. Under the latter model, one error rate for each item is estimated whereas in the former model this error rate is assumed to be the same for each item. Because the item-specific error rate model provides more information for the estimation procedure, we assume that the fit of the model to the data is better than the more restrictive Proctor model. To put this differently, we assume to get a smaller chi-square for the former model at a cost of losing degrees of freedom. This drop in chi-square of the less restrictive item-specific error rate model relative to the loss of degrees of freedom can easily

be tested for statistical significance.[4]

However, researchers also compare the fit of models that are not nested or hierarchically related. The choice between these models must be based on a comparison of indices of fit or other criteria because non-hierarchical models cannot be compared statistically. In these circumstances, often the chi-square statistic is used to determine the fit of a latent class to the data. For instance, the choice between the item-specific error rate model and the type-specific model cannot be statistically determined because these models are not subsets of each other. Instead, the chi-square values of these two models may be compared so that the model with the lowest chi-square value relative to the degrees of freedom suggest the best fit to the data.

The hierarchical relations between the latent class models can be summarized as follows:

-Proctor's model and Goodman's model both are hierarchical with respect to Guttman's model.

-The Dayton-Macready hybrid models (plus unscalable class), and Goodman's model are hierarchically with respect to Guttman's model.

-The item-specific error rate model, the type-specific model, and the false-positive/false-negative error models are hierarchical with respect to Proctor's model.

-The latent distance model is hierarchically related to the

---

[4]In the case of hierarchical related models, we compare the likelihood-ratio statistic L^2 instead of the Pearson χ^2 statistic since the former can be partitioned exactly (Fienberg, 1983:58).

item-specific error rate model.

-The Dayton-Macready hybrid models (plus unscalable class) are hierarchical with respect to each of the response error models.

4.6 Multi-Group Latent Class Model

The scaling models considered here are multiple-group generalizations of the preceding scaling models. These models enable us to compare statistically scaling models from two or more groups by a simultaneous analysis of the latent structure of the data for the groups in question. The procedure for simultaneous latent structure analysis is described by Clogg and Goodman (1984; 1985; 1986) and the interested reader is referred to these sources for the technical details. As for the single-group latent class models, it is assumed that the latent variable is discrete and no assumptions are made about the distribution of the latent variable.

In order to compare the differences between parameter values across groups, we first test whether the same type of model is valid for more than one group, separately. If so, we will test which parameters are constant across groups by placing various restrictions on the parameters of the multiple-group models. Clogg and Goodman (1984, 1985, 1986) discuss three basic types of restrictions: within-group restrictions, across-group restrictions and combinations of the two. Within-group restrictions refer to any kind of restriction imposed on the parameters of the model and are referred to as simply *restrictions*. The models discussed in

the present study are all restrictive models. The across-group restrictions refer to equality restrictions of the corresponding parameters across the groups and are termed *homogeneity constraints*. If there are no across-group restrictions on the parameters (i.e., if we assume that all parameters are different for the groups in question), a model of complete *heterogeneity* is obtained.

For example, assume that we generalize Proctor's single-group model to a multiple-group model. First, we place the same restrictions on the conditional probabilities of each group, separately, as discussed above. This means that the parameters of each model are "restricted". If we next allow the error rate *a* to vary across groups, the model is considered as heterogeneous in the response error rate. However, if we constrain the error response parameter *a* to be equal across groups (i.e., the items are equally reliable indicators of the latent variable in each group), this model assumes homogeneity of the response error rate. Note that this latter model is a restricted model with a homogeneous latent structure.

In the present study, we will compare restricted-heterogeneous models with restricted-homogeneous models. This can easily be done by checking whether the differences between the chi-squared values of the heterogenous models differ significantly from models with inter-group homogeneity constraints because the latter are hierarchically related to the former.

5. LATENT TRAIT MODELS

5.1 Introduction

In Chapter 3 we introduced two subclasses of latent structure models: the latent class and the latent trait models. In the previous chapter, we discussed scaling models that are all special cases of latent class analysis. Here we will introduce the mathematical models that define the various latent trait models.

Latent trait models and latent class methods used to construct scaling models have some important similarities and differences. First, both models assume a one-dimensional ordering of the items and the classes of the latent variable. Second, latent class models as well as latent trait models assume a *probabilistic* relationship between a person's response given her position on the latent variable. In contrast to the deterministic Guttman model, these models take into account the fact that we can never be sure a respondent will answer an item positively, whatever her position on the latent variable. Third, both types of models share in common their reliance on the principle of local independence; i.e., item responses are conditionally independent, given the same value on the latent variable or in the same latent class. As a consequence, in both models the probability of responding positively to an item is independent of the other items included in the scale. Fourth, in both analyses model-data fit can be assessed by comparing the observed and expected patterns of response frequencies, in other words, the models are falsifiable. In

Chapter 2, this feature was referred to as representational measurement.

Despite these similarities, there is one key difference between the two approaches. Latent class models assume a discrete distribution of the latent variable whereas latent trait models view respondents as being distributed continuously over the underlying latent variable. Therefore, the models differ in their specification of the relationship between the latent variable and the probability of a (positive) response. In latent class models, both items and persons are assigned to one of the latent classes, and in this process, estimates are obtained of the proportion of respondents who are (mis)classified based on the best fit to the data. Latent trait models specify the relationship between a person's position on the latent trait, the item and the observable response their interaction is supposed to produce. Thus, rather than estimating proportions of respondents in a certain class directly, latent trait models describe the probability of a response as a function of a set of model parameters that represent locations of persons and items along a continuum.

In the present study, we consider both non-parametric as well as parametric latent trait models. These models assume that a respondent's chance for a positive answer increases with her value on the latent variable but decreases with item difficulty. However, in the non-parametric Mokken model no assumptions are made about the functional form of the ICC's. It requires only that the item trace lines are monotonically increasing and that they

do not intersect. If the model fits the data, the resulting scale scores and item difficulties constitute ordinal values. In contrast, the parametric models define explicitly the functional form of the ICC's and, as a consequence, we can make measurements on an interval level.

In the next sections, we will first discuss Mokken's method of scale analysis. After that, that we will present three parametric latent trait models, usually referred to as, the one-, two- and three-parameter models. Finally, a comparison between the two approaches will be made.

5.2 The Non-Parametric Mokken Model

The Mokken model is a stochastic elaboration of Guttman's scale analysis (Mokken and Lewis, 1982).[1] It is applied to dichotomous items for which one or the other is designated as "positive" with respect to the latent variable of interest (here attitude). The model treats the attitude as a single latent trait on which the person's location is represented by the parameter θ and the item's location is represented by the parameter δ. Given a reasonably unidimensional set of items, that is, one dominated by the latent variable measured, the person parameter θ can be estimated by the number of items to which a person responds positively, and the item parameter (δ) can be estimated by the proportion of people who responds positively. The former is referred to as the person's scale score and the latter as

---

[1]Note that both Andrich (1985) and Schwartz (1986) argue that the Rasch model is the only probabilistic analog of the Guttman model. This belief overlooks the development of Mokken scaling in the early seventies (Mokken, 1971).

the item difficulty.

5.2.1 The Assumption of Double Monotony

The Mokken model specifies the relation between the item and latent trait in terms of an item characteristic curve (or ICC) denoted by $P(X_i|\theta)$. As the formal expression indicates, this curve represents the probability of a positive response on an item i, given respondent j's location θ on the latent trait. An important feature of the Mokken model is that unlike the parametric latent trait models (see Section 5.3), it makes no assumption about the functional form of the ICC. For this reason, we refer to the Mokken model as non-parametric, and the resulting scale scores and item difficulties constitute ordinal, rather than interval or ratio values.

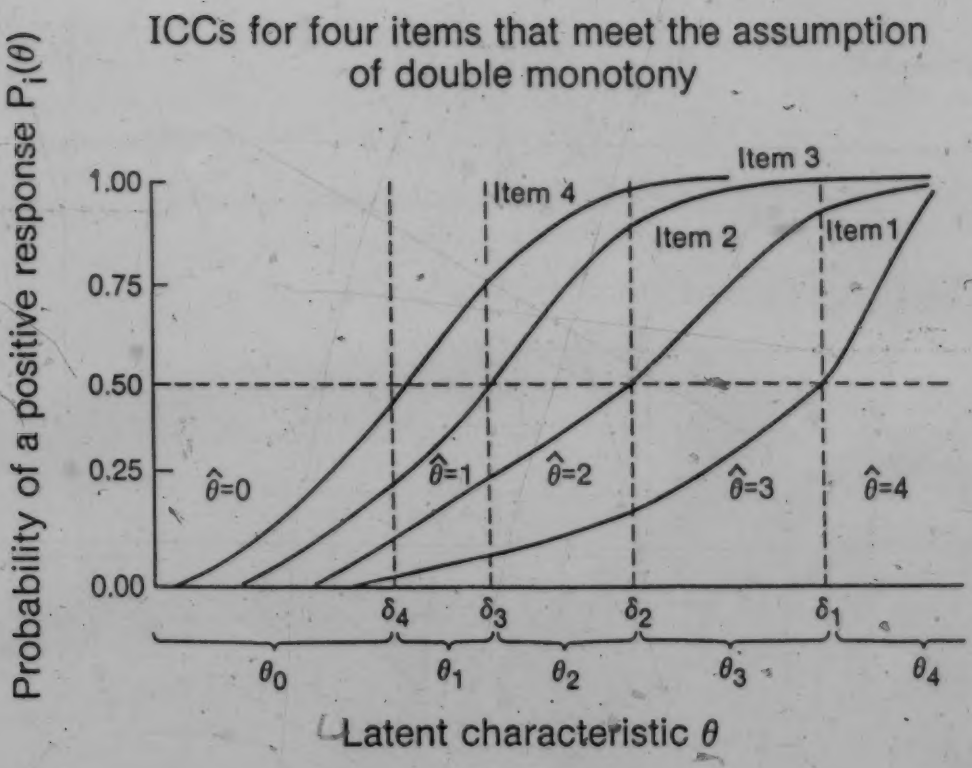
Instead, the only constraint the Mokken model puts on the ICC's is referred to as the assumption of double monotony. The first requirement of this assumption is that for any item in a Mokken scale, the probability of positive response increases as θ increases. To put this more formally, for any two persons i and j, where θ_i is less than θ_j the probability of a positive response on any item in the scale is less for person i. This requirement is referred to as monotone homogeneity (Mokken and Lewis, 1982) and is a necessary requirement for unidimensional measurement of respondents.

The other requirement is that for any value of θ, the probability of a positive response *decreases* with difficulty of the item. This means that the order of item difficulties

remains invariant over the values of θ, or put graphically, the ICC's do not intersect. Given these two requirements of the assumption of double monotony, it becomes possible to define unambiguously the difficulty of an item as the θ of a person who responds positively to the item with a probability of 0.5.

Figure 5.1 graphically illustrates the properties of a Mokken scale. It contains the ICC's of four different items. Item 4 is the easiest, followed by items 3, 2, and 1 in order of increasing difficulty. The items satisfy the assumption of double monotony; the ICC for each item increases with θ and none of the ICC's intersect. Note that the value of δ for each item is found by drawing a line from the ICC to the θ axis at the point on the ICC where the probability of a positive response is 0.5. Also note that the ordinal scale scores are defined in terms of the (unobserved) values of θ. Finally, note that, within the constraint of double monotony, the functional form of the ICC's differ and that the ICC's do not intersect.

FIGURE 5.1


ICCs for four items that meet the assumption of double monotony
Probability of a positive response $P_i(\theta)$
1.00
0.75
0.50
0.25
0.00
Item 3
Item 4
Item 2
Item 1
$\hat{\theta}$=0
$\hat{\theta}$=1
$\hat{\theta}$=2
$\hat{\theta}$=3
$\hat{\theta}$=4
δ_4
δ_3
δ_2
δ_1
θ_0
θ_1
θ_2
θ_3
θ_4
Latent characteristic θ

5.2.2 Coefficients of Scalability

To test the requirement of monotone homogeneity Mokken developed three related coefficients of scalability which are based on the assumption of local independence. The first, $H_{i,j}$, measures the homogeneity or association between each pair of items. The second, H_i, measures the homogeneity of a particular item with respect to all other items and is obtained by aggregating across the coefficients of the relevant item pairs. The third, H, measures the homogeneity of the scale as a whole by aggregating across the coefficients for the individual items.

According to Loevinger (1948), the coefficient for the homogeneity of an item pair essentially measures the association in the two by two table that is obtained by cross classifying the two items. Table 5.1 illustrates such a table. In constructing this table, we assume that item *i* defines the rows, and item *j* defines the columns, and that item *i* is more difficult than *j*.

Table 5.1: The Cross-tabulation of Two Items

	Response to Item *j*		
Response to item *i*	1	0	Row Total
1	f(1,1)	f(1,0)	f(1,.)
0	f(0,1)	f(0,0)	f(0,.)
Column Total	f(.,1)	f(.,0)	f(.,.)

Item *i* is assumed to be more difficult than item *j*.
"1" denotes a positive response; "0" denotes a negative response.

Under Guttman's deterministic model we would expect the top right-hand cell, the error cell, to be empty; i.e., f(1,0)=0. Under the model of statistical independence (or no association between *i* and *j*), we would expect the frequency of the error cell to equal the product of the marginal frequencies divided by the sample size; i.e., e(1,0) would equal f(1,.)f(.,0)/f(.,.). As given in (1), H_{ij}, the index of item pair homogeneity measures the proportional difference between cell frequency of the error cell expected under independence and the actual cell frequency.

$$H_{ij} = \frac{[\, e\,(1,0) - f\,(1,0)\,]}{e\,(1,0)} \qquad (1)$$

where e(1,0)=f(1,.)f(.,0)/f(.,.).

Readers familiar with the convention of using the letters *a*, *b*, *c*, and *d* to represent the cell frequencies of a 2 x 2 table may find the following formula for H_{ij} more convenient.

$$H_{ij} = (ad - bc) / (a + b) (b + d) \qquad (2)$$

They also may notice the similarity between this index and other measures of association for 2 x2 tables. When items

are independent, H_{ij} will be zero; when the error cell is empty, H_{ij} will equal unity.

The coefficient of item homogeneity, H_i, is given in (3). It simply aggregates the observed and expected frequencies for the error cell for each 2 x 2 table that cross-classifies item i with the other items in the scale. H_i is analogous to the item-total correlation used in reliability analyses (Nunally, 1978). It will be zero when item *i* is independent of the other items used in the scale. It will attain a maximum value of unity when the error cell of the relevant 2 x 2 tables is empty.

$$H_i = \frac{\sum_{j=1}^{k} e_{ij} - \sum_{j=1}^{k} f_{ij}}{\sum_{j=1}^{k} e_{ij}} \qquad (3)$$

where $i \neq j$.

In terms of formula (2), the numerator of (3) consists of the sum of the differences between the diagonal and off-diagonal cross-product terms for the two by two table. The denominator of (3) sums the product of the appropiate marginals of these tables.

The coefficient of scale homogeneity, H, is given in (4). It aggregates the observed and expected frequencies used to calculate H_{ij} for *all* item pairs.

$$H = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} e_{ij} - \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} f_{ij}}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} e_{ij}} \qquad (4)$$

where $l=j+1$.

Again, in terms of formula (2), the numerator of (4) consists of the sum of the differences between the diagonal and off-diagonal cross-product terms for *all* two by two tables; the denominator (3) sums the products of the appropriate marginals from these tables. The coefficient is analogous to measures of scale reliability, e.g., Cronbach's alpha (Nunally, 1978), except that a particular cross-classification of an item pair is used only once in the calculation of H rather than twice as is the case in the calculation of Cronbach's alpha. H will be zero when all the items are mutually independent and will attain a maximum value of unity when the error cells for all the two by two tables are empty.

Sample estimates $\hat{H}$ and $\hat{H}_1$, can be obtained by inserting the sample relative frequencies f(1,.)/f(.,.) and f (1,1) in equation (3) and (4). Asymptotic sampling theory for these estimates is completely developed by Mokken (1971: 157-160), and includes the following results:

1. One-sided tests for a scale ($H=0$ vs. $H>0$) and for individual items ($H_i=0$ vs. $H_i>0$);
2. Confidence intervals for H and H_i; and
3. Tests of equality of H (and H_i) for different populations.

It may therefore be concluded that H seems to satisfy fully four prerequisite criteria for a coefficient of scalability proposed originally by White and Saltz (1957:82), together with a fifth one extending its usefulness:

1. Its theoretical maximum is 1 and hence invariant over scales.
2. Its theoretical minimum is 0, assuming monotone homogeneity and hence is invariant over scales.
3. It is possible to evaluate scales as a whole with H and also to evaluate the scalability of individual items with the item coefficient H_i.
4. It is possible to test theoretically interesting hypotheses about H and H_i.
5. It is possible to construct approximate confidence intervals for H and H_i.

The coefficients of scale and item homogeneity allow the researcher to judge the scale (or scalable set) as a whole and the scalability of individual items. Mokken (1971) has established a set of criteria for using all three coefficients to judge the homogeneity of a scale. First, all the H_{ij} should be greater than zero. Second, all the H_i and, therefore, H should be greater than a predetermined constant (c). The following classification of scales was suggested:

.50 ≤ H: a strong scale;

.40 ≤ H < .50: a medium scale; and

.30 ≤ H < .40: a weak scale.

In practice, when an item does not meet these criteria it is eliminated from the scale. The coefficients H_i and H are then recomputed for the remaining *subset* of items, the coefficients are checked against the criteria (e.g., $c \geq .30$), and the process is repeated until a sufficiently strong Mokken scale is obtained. The concept of a strong scale corresponds to the original strong requirements for the Guttman scale, with values near unity indicating nearly perfect scales.

5.2.3 Test of Double Monotony

The coefficients of scalability, H and H_i, and the definition of scalability are used to test the monotone homogeneity of the items. To test the assumption of *double* monotony additional criteria are introduced to test the assumption that the ICC's do not intersect.

This test involves an inspection of the P and P_o matrices which contain the probabilities of two positive and two negative responses, respectively, to all possible pairs of items. According to the assumption of double monotony, item responses are conditionally independent, given the same value of θ, so that the (conditional) probability of joint response for persons with the same value of θ equals the product of their marginal probabilities of response. In other words, the test of double monotony is based on the

assumption of local independence which can be empirically verified by the observable unconditional probabilities of each item pair in the P and P_o matrices. Since marginal independence is assumed under the assumption of a given θ value, the unobservable conditional probabilities imply the observable unconditional probabilities.

Thus, when items 1, 2 and 3 represent decreasing levels of difficulty, the (conditional) probability of a pair of positive responses will be greatest for items 2 and 3, followed by items 1 and 3, and 1 and 2. Similarly, the (conditional) probability of a pair of negative responses will be greatest for 1 and 2, followed by 1 and 3, and 2 and 3.

The test of double monotony, then, involves an inspection of P and P_o matrices. When the rows and columns are ordered from top to bottom and from left to right according to decreasing levels of item difficulty, the probability of a pair of positive responses should increase in the P matrix, while the probability of a pair of negative responses should decrease in the P_o matrix. Items that do not fit the pattern are removed from the scale. The statistical significance of departures from this pattern can be tested by means of a simple run test (Molenaar, 1983; Siegel, 1956).

5.2.4 Cross-Population Comparisons of the Mokken Test

Once the researcher finds a scale, she can test the robustness of the scale across different populations. Essentially, this test involves comparison of the values of

H for different samples against an average or pooled value of H. Mokken has developed the test statistic T for testing the null hypothesis that the individual values of the homogeneity index, H_b, for subpopulations (b=1, 2, 3, ..., p) are equal. Given in equation (5), this statistic is asymptotically distributed as a chi-square with b-1 degrees of freedom (where b is the number of groups) when the null hypothesis is true. Thus, if the researcher accepts the null hypothesis, she can pool the groups and, in effect, use the mean value of H_b, $\bar{H}$, to estimate a single value for the scale homogeneity that applies to all groups.

$$T = \sum_{b=1}^{p} \frac{(H_b - \bar{H})^2}{S^2(H_b)} \qquad (5)$$

where S^2 denotes the variance of the distribution of H. (See Niemoller, et al., 1982; Kingma and Reuvekamp, 1986b for additional discussion of the test).

In addition, the Mokken Test provides a statistical basis for deciding whether the ordering of a given set of items is invariant across different subpopulations. Molenaar (1982) does this by dividing his sample into two or more strata using a variety of stratifying variables, such as intermediate values of the respondent's scale score. If the assumption of double monotony holds for a population, the order of the item difficulties should remain invariant

(subject to sampling error) for any sample of the population, and if the model holds for a population of items, the order of the respondents (based on scale scores) will remain invariant for different samples of items (again subject to sampling error). This feature is referred to as "specific objectivity" (Rasch, 1966), a concept used in connection with the Rasch model discussed below.

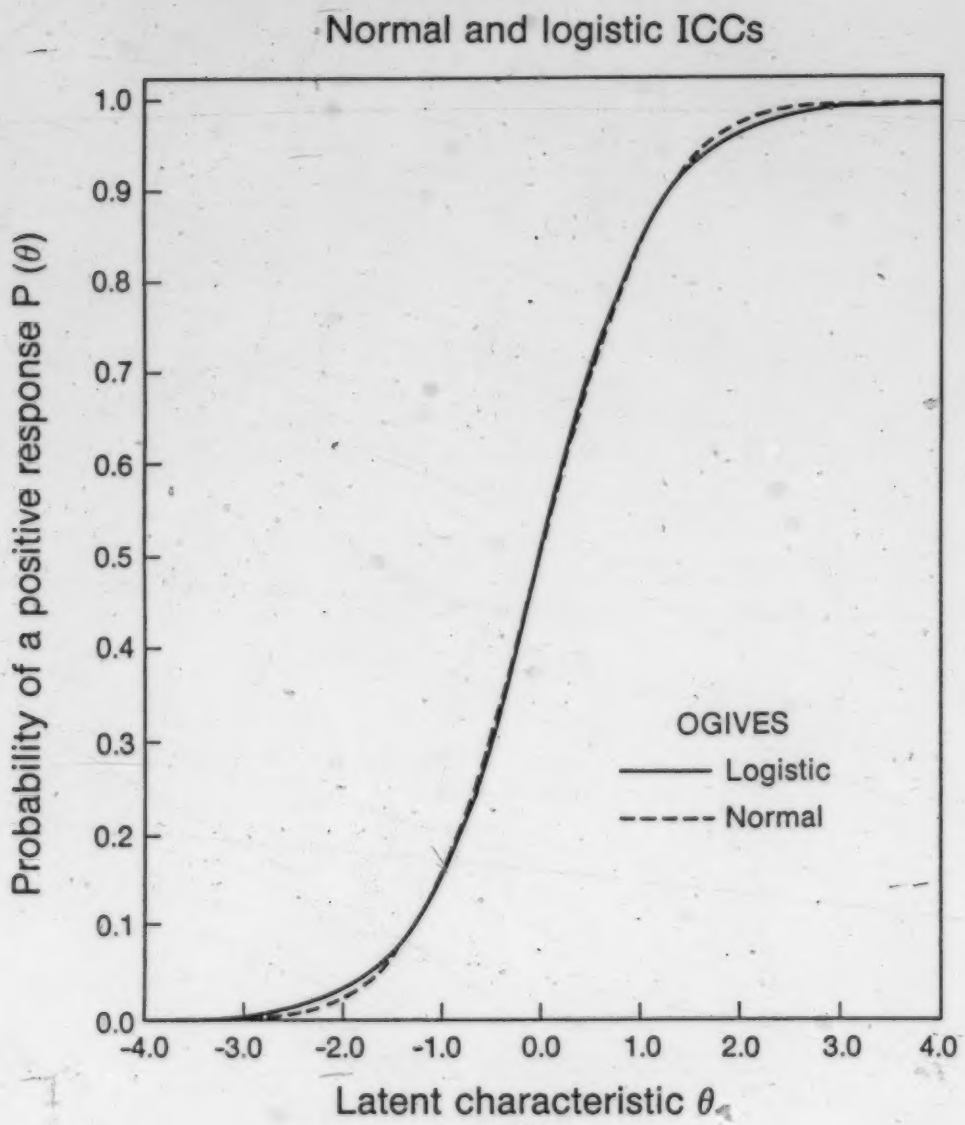
5.3 Parametric Models: Normal versus Logistic Ogive

In sub-section 3.3.1, in which parametric latent trait models were discussed, the distinction was made between the normal ogive model and the logistic ogive models. In this section the item characteristic curves (ICC) of these models will be discussed.

The logistic ogive model is a function which nearly coincides with the normal ogive model (see Figure, 5.2). The logistic ogive can be written as:

$$\Psi = \exp(x)/(1+\exp(x)) = 1/(1+\exp(-x)) \quad (6)$$

FIGURE 5.2


Normal and logistic ICCs
Probability of a positive response P (θ)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Latent characteristic θ
OGIVES
Logistic
Normal

Haley (1952), quoted in Birnbaum (1968), showed that:

$$|\Phi(x)-\Psi[(1.7x)]|<0.01 \quad \text{for all } x.$$

In words, the normal ogive $\Phi(x)$ and the logistic ogive $\Psi(x)$ with standard deviation 1.7 do not differ more than 0.01 uniformly for all x. Molenaar (1974) showed that indeed fixing the scaling factor D to 1.7 minimizes the maximum difference between the two cumulative curves. Thus to maximize agreement between the two functions, we usually take D = 1.7. Since the logistic ogive is mathematically far simpler than the normal ogive, the first one can be used as an approximation of the second one.

The popular models for unidimensional latent trait models are the one, two, and three-parameter logistic models. Although these models have been developed primarily in connection with ability testing, present research shows that they might be applied to other kinds of latent traits as well, such as attitudes. The number of parameters required to describe an ICC will depend on the particular latent trait model. Three parameter models are distinguished according to the number of parameters: one, two, or three.

5.3.1 Two-Parameter Logistic Model

Birnbaum (1968) proposed a latent trait model in which the ICC takes the form of a two-parameter logistic function:

$$P_i(\theta)=\{1+\exp[-1.7a_i(\theta-b_i)]\}^{-1} \qquad (7)$$

where:

$P(\theta_i)$ is the probability that a respondent with latent trait score θ answers item i correctly,

a_i is the discrimination parameter,

b_i is the item difficulty.

With the two-parameter logistic model, ICCs vary in slopes as well as in item difficulty. That is, some items discriminate between respondents with high and low scale scores better than others.

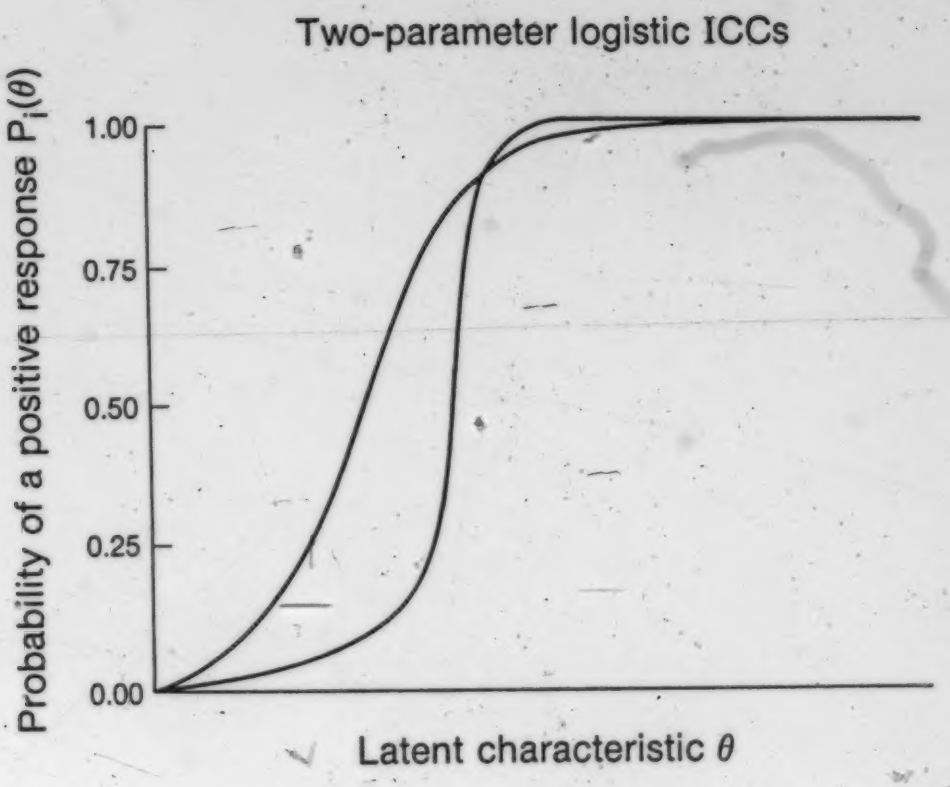
Figure 5.3 shows two ICCs for the two parameter model. The parameter b_i represents the point along the θ continuum at which respondents have a 50 percent chance of responding positively. Parameter b_i is a location parameter; the more difficult the item, the further the curve is to the right. Parameter a_i controls the steepness of the ICC and is proportional to the slope of $P_i(\theta)$ at the point $\theta=b$.* If the scaling factor D is set equal to 1.702, it can be shown that the slope of the ICC is 42.55 percent of a_i at $\theta=b_i$ (Lord, 1980:13). Note that the ICCs for this model are allowed to intersect and that in that case the set of items is not double monotone (see Section 5.2.1).

A consequence of intersecting ICCs is that a respondent's θ value depends upon the particular pattern of item responses rather than on the total score (=total number of items endorsed). Thus, under this model respondents with the *same* total score could obtain different θ values if they

---

*Note that the letters a and b mean the opposite to their use in linear regression where the former refers to the intercept and the latter to the slope of the regression line.

FIGURE 5.3


Two-parameter logistic ICCs
Probability of a positive response $P_i(\theta)$
1.00
0.75
0.50
0.25
0.00
Latent characteristic θ

answered different items positively. In practice, this means that the researcher cannot reproduce a respondent's response pattern by means of her total score, a feature that might contradict the widespread belief in the fundamental meaningfulness of an unweighted sumscore; i.e, respondents with the same sumscore should receive the same θ value.

5.3.2 Three-Parameter Logistic Model

The three-parameter logistic model (Birnbaum, 1968) is the most general of the three logistic models. This model can be obtained by adding a third parameter, denoted c_i, to the two-parameter model. The mathematical form of the three-parameter logistic curve can be stated as follows:

$$P_i(\theta)=c_i+(1-c_i)\{1+\exp[-1.7a_i(\theta-b_i)]\}^{-1} \qquad (8)$$

where c_i is interpreted as the guessing parameter in ability testing research. The three-parameter logistic model allows differences between the slopes of the ICCs *and* lower asymptote. This last feature of the ICC is particularly appropiate when respondents with low standings on the latent trait can occasionally respond correctly to difficult items. However, the "guessing parameter" is developed in the context of ability testing and does not apply to attitude measurement for *dichotomously* scored items (as in the present study), since it seems unlikely that respondents

really "guess" the answers to the items.'

Figure 5.4 provides two three-parameter ICCs. Note that the probability of guessing among very low respondents is assumed to be higher for item 1. When there is guessing, the b_i corresponds to the value of θ at the inflection point of the ICC; that is, b is the ability level where the probability of a positive answer is halfway between c and 1.0. or $(1+c/2)$. In Figure 5.4, item 1 is endorsed less frequently than item 2 and therefore has a larger b parameter than item 1. Parameter a_i of the three-parameter logistic model is $(1-c_i)(D/4)a_i$ when evaluated at $\theta=b_i$ and, as a consequence, the slope depends upon both a_i and c_i. Note that the larger the c parameter the lower the steepness of the ICC, that is, an item with a large c parameter distinguishes less clearly between those respondents with high and low values on the latent trait than an item with a lower c parameter.'[10]

5.3.3 One-Parameter Logistic Model

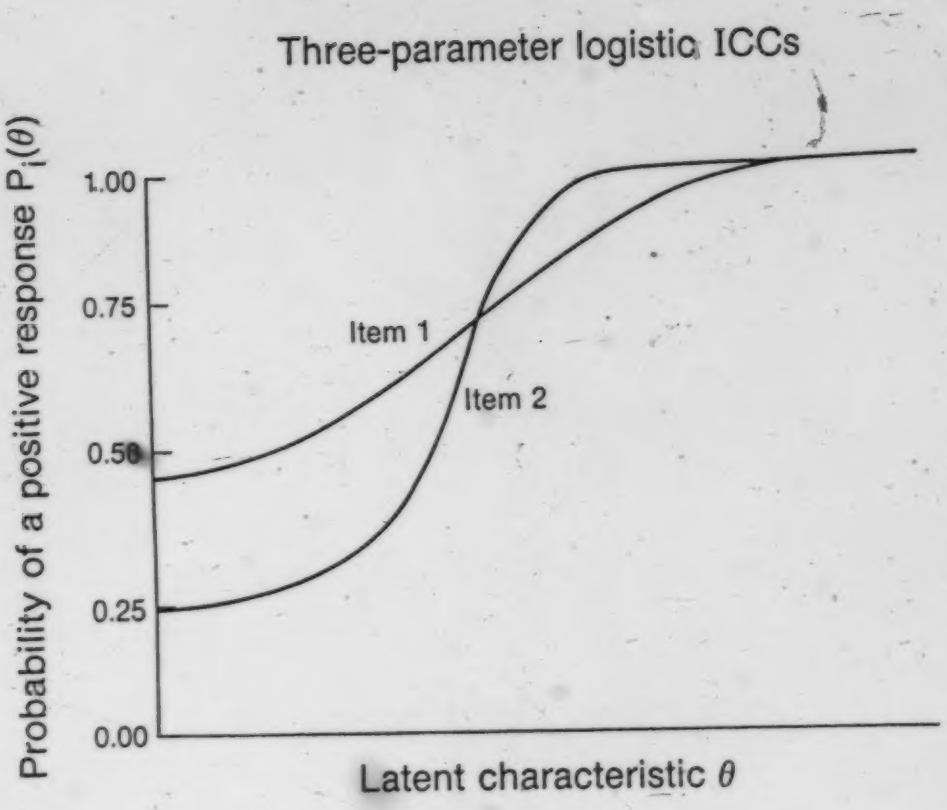
Finally, the one-parameter or Rasch model, named after the Danish mathematician (Rasch, 1966), was developed independently of other latent trait models and along quite different lines (Hambleton and Cook, 1977). This model is a

---

'If an acquiscent (or some other) response set occurs, one could interpret this third parameter as a measure of the extent to which the item is subject to this response set.

[10] Barton and Lord (1981) describe a four-parameter logistic model. In this model it is assumed that respondents with a high value on the latent variable may not respond positively to an item, due to lack of interest, carelessness etc. The authors, however, were unable to find any practical gains that accrued from the use of the model.

FIGURE 5.4


Three-parameter logistic ICCs
Probability of a positive response $P_i(\theta)$
1.00
0.75
0.50
0.25
0.00
Item 1
Item 2
Latent characteristic θ

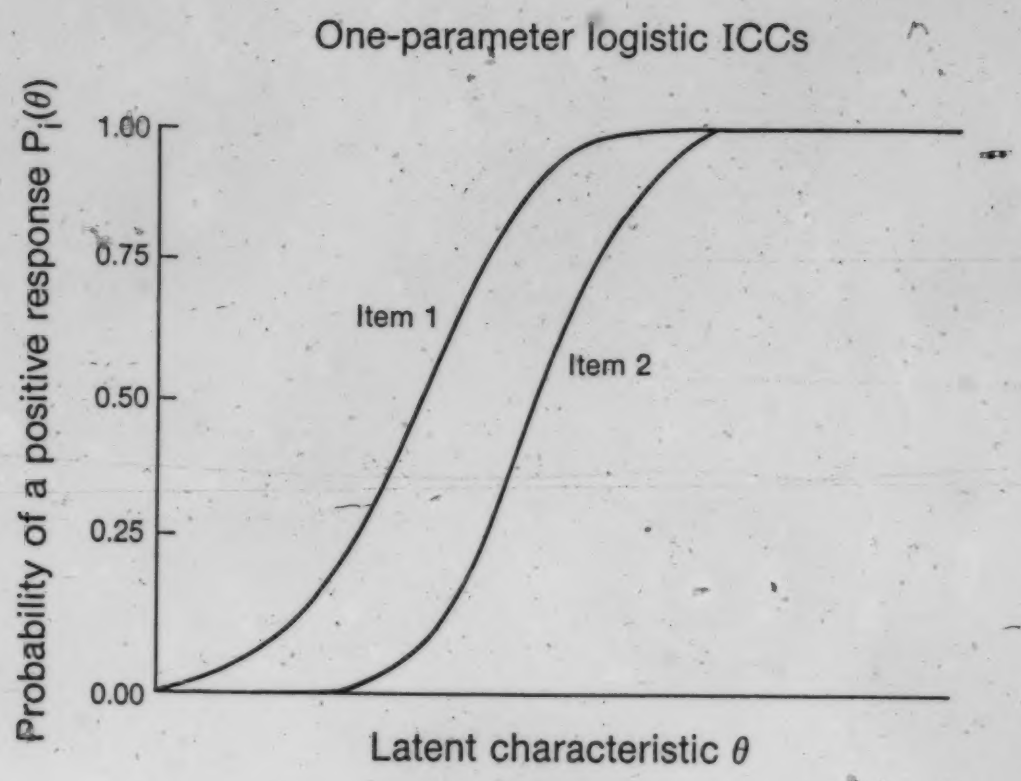
special case of the three-parameter logistic model that results from setting $c_i=0$ and $a_i=1$ for all items. In words, first, all items are assumed to have equal discriminating power, and, second the c parameter is assumed to be zero. Items are allowed to vary only in terms of their difficulty. The form of the one-parameter logistic model is:

$$P_i(\theta)=1+\exp[-1.7(\theta-b_i)]^{-1} \qquad (9)$$

Figure 5.5 represents two one-parameter ICCs. It is easy to see that the tracelines parallel each other, that is, each item has a discrimination parameter which is fixed at a value of 1 for all the items in the set. As a consequence, respondents with the same sumscore will obtain the same θ-value. As in the two-parameter logistic model, the b_i parameter is the point along the θ continuum where the probability of a positive response is .5.

The one-parameter logistic model has many desirable features. The Rasch model is the only logistic model that has *sufficient statistics* for person and item parameters (Andersen, 1973). This means that the item totals are sufficient statistics for the item parameters and that the number of positively responded items is a sufficient statistic for the person parameter. Thus the pattern of positive-negative responses does not provide any additional information.

FIGURE 5.5


One-parameter logistic ICCs
Probability of a positive response $P_i(\theta)$
1.00
0.75
0.50
0.25
0.00
Item 1
Item 2
Latent characteristic θ

This statistic t is given by

$$t=x_1+....+x_k$$

where $x_j=1$ if the respondent answers the item positively and $x_j=0$ if a negative answer is given. In other words, the value of the sufficient statistic exhausts the information contained in the data about the parameter. It is easy to show that, for instance, in the two-parameter model there is no sufficient statistic for the person parameter, because the *weighted* sumscore contains the unobservable item discrimination parameter b_i, and so is itself unobservable.

$$t=b_1x_1+.....+b_kx_k$$

As a result, there is no sufficient statistic for the person parameter, and hence no way to eliminate the person parameters from the estimation of the item parameters. (Fischer, 1974; Wright and Masters, 1982; Schwartz, 1986). No sufficient statistic exists for the three-parameter model (Swaminathan, 1983).''

Another property that follows from the Rasch model is that of *"specific objectivity"* (Rasch, 1966; Fischer,

---

'' In the case of latent class analysis, the number of positive responses is not a sufficient statistic for class membership. Here the pattern of responses is of eminent importance and the frequencies of occurence of these observed response patterns serve as the sufficient statistics (Dayton and MacReady, 1980).

1974:ch.19). Specific objectivity is a philosophical concept that is based on Rasch's view that science is a matter of comparison. According to Rasch, scientific statements in social sciences deal with comparisons, and the comparisons should be objective, i.e., independent of the instrument used. In social sciences, this means that the comparison between two items should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other items within the considered set of elements were or might also have been compared.

Symmetrically, a comparison between two individuals should be independent of the items used for the comparison; and it should also be independent of which other individuals were also compared, on the same or on other occasions (Rasch, 1961:351-352).'[3]

Specific objectivity and estimation sufficiency are two sides of the same approach to inference; i.e., that the statistical model on which inference is to be based be seperable (factorable) in its parameters. As Andersen (1973) shows, since the Rasch model has the feature of sufficient statistics, we can estimate the parameters separately by conditioning the likelihood of the data on the marginal totals (the sufficient statistics) for the parameters to be eliminated. This method of parameter estimation is referred to as the Conditional Maximum Likelihood (CML) estimation and is also used in the log-linear representations of the

---

'[3]Note that Thurstone (1927), already required measuring instruments which are independent of the object of measurement. (see also Section 2.5.1).

model (See, Clogg, 1987; Duncan, 1984a). We will discuss parameter estimation in Chapter 6 along with the available computer programs.'[13]

Further, since the model is the simplest of the logistic models, it is the easiest model to work with. Finally, since the model involves fewer item parameters the problems with parameter estimation are considerably smaller than for the more general models (see Chapter 5).

However, the simple form of this model may be its chief weakness when applied to empirical data. The assumption that all item discrimination parameters are equal is very restrictive and will only be appropriate for carefully pretested and selected sets of items. Therefore, the non-parametric Mokken model might provide a useful tool for researchers who want to avoid the strong parametric assumptions of the Rasch model or as a method for selecting items that may fit the Rasch model.

5.3.4 The Rasch versus Mokken Model

When we investigate the general properties of the Rasch model and the Mokken model, it can be shown that the former model is a special case of the latter. Both models assume double monotony which implies in the case of the Mokken model that the *ordering* of the items is specifically objective: in any group of subjects item *i* is more difficult

---

[13] Wright and Panchapakesan (1969), Wright and Masters (1981) refer to this principle as "sample free estimation", whereas Andersen uses the terminology "estimation of person (or item) parameters independent of the actual values of the item (or person) parameters" (See e.g. Andersen, 1973).

than item *j* (Molenaar, 1982). In the parametric Rasch model this property corresponds to the stronger requirement that the difference between item difficulties must be invariant across samples. Because both models assume double monotony of the ICCs (i.e., the ICCs are not allowed to intersect), each individual's response pattern can be predicted with a high degree of certainty from her scale score. As Mokken (1971) points out, the number of items a respondent endorses provides the best estimate of the person's scale score in the Mokken model. In the parametric Rasch model also, the total score serves as the best estimate of the person's scale score on the latent variable, the θ parameter, because as a sufficient statistic the sum score contain all the (probabilistic) information concerning θ. Thus in both models, the item responses are stochastically reproducible from the sumscores with the Guttman pattern most likely.

Further, in cross-national comparative research (the concern of the present study), the strong parametric measurements are of importance due to the invariance of item and person parameter estimation across subpopulations. In addition, a Rasch scale is an interval scale which enables the researcher to compare differences in scale scores between populations. With respect to the non-parametric Mokken model, the Mokken Test provides a method for reducing a set of scalable items to a subset that is robust across populations. Since the Mokken model assumes only an ordinal scale, the distance (intervals) that separates the scale score of the items plays no role in the outcome of the Mokken Test. The test is sensitive to cross-population

reversals in item ordering and differences in the homogeneity between the item and the rest of the scale.

In sum, the parametric Rasch model has one major advantage over the Mokken model; the higher level of measurement. This gain of information is bought at a certain price since the very strong assumptions of the Rasch model are hard to meet in sociological research where we have only a low level of knowledge concerning items and latent variables. Therefore:

> "it seems legitimate to try to find starting points for scaling models which do not rely too heavily upon specific parametric assumptions, as these lead to procedures of inference and estimation that are too pretentious and intricate for the level of information and the precision that can be claimed for the data used in actual measurement. At the same time we may want these models and procedures to bear a clear relationship to the parametric models, although the former are less specific in their assumptions and more in harmony with our limited knowledge concerning the data. If we use these simpler but related models primarily for investigating the scalability of sets of items, their relation to more sophisticated parametric models may enable us to specialize them to a specific parametric model in the more advanced stages of the research" (Mokken, 1971:173).

In the present study we will start with the Mokken method as a preliminary step in the development of a cross-national scale for the abortion items. If a particular set of items meet the Mokken criteria for a robust scale, we will apply more restrictive techniques such as Rasch's model or McDonald's harmonic analysis.

5.4 Latent Class Models and the Assumption of Double Monotony

As noted before, both the Rasch model and the Mokken model assume the assumption of double monotony; i.e., 1) the probability of a positive response is a decreasing function of the item difficulty, and 2) an increasing function of a person's position on the latent trait.

In terms of latent class models, the first requirement of this assumption implies the following inequality constraints on the conditional probabilities of the item specific-error rate model (see Section 4.4.1.2): $a < b < c < d$. Note further, that the type-specific error rate model (see Section 4.4.1.3) is inconsistent with the assumption of double monotony. When we apply this assumption to the type-specific error model it implies that $a < b < c < d$ but also that $1-b < 1-c < 1-d < 1-e$. Clearly, both assumptions cannot be true.

A main problem with the application of latent class models from the perspective of the assumption of double monotony is that the existing software for latent class models does not permit the specification of inequality constraints on the estimates of the error rates. This limitation makes a comparison between the results of the latent class models and the latent trait models more complicated.

5.5 Assigning a Metric to an Ordinal Scale

A fundamental problem in social research is the often implicitly accepted assumption of invariance of the relationship between the indicator and latent variable across subsamples in, for instance, regression and factor analysis models. These techniques claim that the slope of the regression line that links the latent variable and the indicator is the same in each subsample; i.e., there is no difference in the unmeasured mechanism that produces a certain score on an item.

However, there may be some reason that for a given measure (item) the slopes are not equal and, as a consequence, general interpretations of the slope will be misleading. For instance, an item may be inappropiate for a culturally different group because it shows a lower correlation with the underlying variable. (In terms of latent trait models, the item has a significantly lower discrimination parameter in that particular group). This difference between groups in slopes may be produced by variables (for instance, "background variables") which are causally related to the latent variable. By not specifying these variables in the model, changes in relationships between the measure (item) and certain other variables could simply represent differences between the relationship of the measure and the latent trait.

In this study we are concerned about a particular aspect of this problem: the development of reference

indicators in latent structure models (Blalock, 1982). This can be explained as follows. Since the latent variable has no metric of its own, we have to select one of the measured indicators as "reference indicator". This reference indicator is a variable whose "loading" or slope on the latent variable (attitude) is set equal to one by the researcher. The term "reference" is used because fixing the loading at one means that the metric of the latent variable will be the same as the metric of the reference indicator. In addition, the loadings of the indicators, which are estimated from the data will also be expressed in terms of the metric of the reference indicator.

In cross-population analysis using LISREL models, the selection of a reference indicator poses a vexing problem since valid comparison requires that the slope of the reference indicator on the latent variable is invariant (Blalock, 1982:80-85).

In contrast to the logistic models, the scales yielded by both the latent class and Mokken method are ordinal, i.e., we do not know the metric of the scale and consequently cannot assume comparability in the metrics of the latent variable.

One of the purposes of the present study is the refinement of an ordinal cross-national scale by assigning a metric to the categories of the scale. We will accomplish this by applying the Rasch model although this model requires assumptions that may difficult to satisfy the data

we use in this study. In addition to the Rasch model, we consider log-multiplicative models discussed by Clogg (1982; 1984a; 1984b) and Goodman (1979b). This method avoids the strong assumption of the Rasch model by assigning metric scores to the ordinal scale values of the cross-national scale based on the (relatively weak) assumption that the latent variable is associated with another "instrumental" variable in a specified matter. In addition, the model enables us to test the assumption of equal intervals of the scale scores across groups by constraining this association to be equal in each population.

In the following we will briefly discuss how these so-called "conditional association" models have been formulated. For a more detailed discussion of these models and applications to social science data see Goodman (1979b), and Clogg (1984a, 1984b).

First, assume that there is a latent variable (attitude) that underlies the responses to a set of items. Second, assume that there is an "instrumental" variable which is associated with the latent variable in a certain form. Third, assume a group variable (e.g. country or sex) and constrain the specified association between the instrumental and the scale scores of the latent variable to be equal across groups. These assumptions lead to the following log-multiplicative association models.

Suppose that a discrete *ordinal* variable with I categories is cross-classified with an *ordinal* variable with

J categories, for each of K groups. Let f_{ijk} denote the observed frequency in the $(i,j,)$ cell of the kth table, and let the corresponding expected frequencies be denoted by

F_{ijk}, for $i=1,....I$; $j=1,....J$; $k=1,....K$.

Let

$$\theta_{ijk}=(F_{ijk}F_{i+1,j+1,k})/F_{i,j+1,k}F_{i+1,j,k}), \qquad (10)$$

for $i=1,....,I-1$, $j=1,....J-1$, $k=1,....K$, denote the basic set of $K(I-1)(J-1)$ conditional odds ratios. This basic set of odds ratios from 2x2 subtables formed from adjacent rows (i.e., rows i and $i+1$) and adjacent colums (i.e., columns j and $j+1$) in a set of K two-way tables contain all the information about the association between the variables and hence an "analysis of association" (ANOAS) approach can be applied to explain the variability in this set of odds ratios.

As Goodman (1979b) and Clogg (1982) show, these associations can be analyzed in a manner analogous to the usual two-way analysis of variance framework by considering models which allow the odds ratios to depend on an overall effect, a row effect, a column effect, and on row and column interaction.

The conditional log-multiplicative assocation model can be described by:

$$\log \theta_{ijk} = \phi_k(\mu_{i+1,k} - \mu_{i,k})(\nu_{j+1,k} - \nu_{j,k}). \quad (11)$$

The quantity ϕ_k describes the overall-assocation between the row (scale scores latent variable) and column variable (scores instrumental variable) for each group. This quantity is multiplied by a factor $(\mu_{i+1,k} - \mu_{i,k})$, which reflects the distance between the rth and the (r+1)th category of the row-variable for each group. The $(\nu_{j+1,k} - \nu_{j,k})$, denote the estimates of the distances between the categories $j+1$ and j of the instrumental variable in the different groups. If the ordering of the row variable is correct, then $\mu_1, \ldots \leq \mu_I$, and the distances $\mu_{i+1} - \mu_i$ will all be nonnegative, with a similar statement applying to the ν_j. The product $(\mu_{i+1,k} - \mu_{i,k})(\nu_{j+1,k} - \nu_{j,k})$ will be nonnegative everywhere, if the correct ordering of categories is used for each variable in each group.

Note that since the $\mu_{i,k}$ and $\nu_{j,k}$ are regarded as model parameters, this product defines a linear-by-linear interaction term (Haberman, 1974).[14] Further, under the assumption that the the model fits the data, the parameters $\mu_{i,k}$ can be viewed as scores (or locations) pertaining to the ith row category for each group and the parameter $\nu_{j,k}$ in this model can be viewed similarly for the jth column category. Thus, dependent upon the constraints of the model,

---

[14]Note that in log-linear models, the $\mu_{i,k}$ and $\nu_{j,k}$ are constants chosen a priori to depict the category scores in question.

either the $\mu_{i,k}$ or the $\nu_{j,k}$ or both can be used to scale the variables in question.

An interesting property of the model is that it is invariant under switches in the categories of the row and/or column categories (Goodman, 1979). This invariance property demonstrates that only the ratios of distances between scores have any importance for the model.

Based on the general model (11), we can define various associations between the two variables. In the following, we present the models that are used here in order to construct an equal interval scale for the ordinal scale scores of the latent variable across groups.

The first model can be written as:

$$\log\theta_{ij(k)}=1 \qquad (12)$$

This model of conditional independence denote that the association between the scale scores of the latent variable (the row variable) and the instrumental variable (the column variable) is independent for each group. Of course, if this model fits the data we cannot estimate the category scores because there is no information available with which to estimate these scores.

The heterogeneous column, row-column effects model is the second model of interest and can be obtained when:

$$\log\theta_{ij(k)} = \phi_k(\mu_{i+1}-\mu_i)(\nu_{j+1,k}-\nu_{j,k}) \quad (13)$$

This model allows the overall association between the row variable and the column variable to differ across groups as well as the association to depend on the level of the instrumental variable $(\nu_{j+1,k}-\nu_{j,k})$ in different ways for the groups. However, the row effects are assumed to be homogeneous under this model, i.e.,

$\hat{\mu}_{i,(1)}=\hat{\mu}_{i(2)}=\hat{\mu}_{i(k)}$ for all i.

The heterogeneous row-column, row-column effect model relaxes the assumption of homogeneous row effects on the association and is obtained when:

$$\log\theta_{ij(k)}=\phi_{(k)}(\mu_{i+1(k)}-\mu_{(k)})(\nu_{j+1(k)}-\nu_{j(k)}) \quad (14)$$

It is easy to see that both the distances between the categories $i+1$ and i of the row variable and the distances between the categories $j+1$ and j of the column variable are allowed to differ across groups. If this model indicates a significant improvement in fit compared to the former model, the scores of the row variable (the latent variable) cannot be considered as the same across the groups.

Finally, the heterogeneous column effect model is used to test the hypothesis of *equal* intervals between the scores of the latent variable. This model can be written as:

$$\log\theta_{ij(k)}=\phi_{(k)}(\nu_{j+1(k)}-\nu_{j(k)}) \quad (15)$$

The model constrains the response categories of the row variable to be equally spaced $(\mu_{i+1(k)}-\mu_{i(k)})=\delta$. If the fit of this model indicates a significant improvement compared to the second model, we accept the hypothesis of equal intervals and, as a consequence, can apply models to this scale that assume equal spacing.

6. DATA AND METHODS

6.1 The Data

The data we analyze in this paper come from the combined files of the General Social Survey (GSS) of Americans conducted by the National Opinion Research Centre (NORC) and the 1982 General Social Survey of West Germans (ALLBUS) conducted by the Zentrum fuer Umfragen, Methoden and Analyses (ZUMA). Both surveys used stratified multistage area probability sampling designs.

The NORC sample universe consists of English speaking persons 18 years of age and older, living in non-institutional arrangements within the continental United States. At the first stage of selection, the primary sampling units (PSU's) employed were Standard Metropolitan Statistical Areas and nonmetropolitan counties selected in NORC's Master Sample. These PSU's were based on figures and other demographic information obtained from 1970 Census Reports. The statistical areas and counties had been stratified by region, or race (or both) before being selected.

The units of selection of the second stage of gathering the NORC sample were census block groups or enumeration districts within Standard Metropolitan Statistical Areas or counties. Before selection, the block groups and enumeration districts were stratified according to region, race and income. Block groups and enumeration districts were then

selected with probabilities proportionate to the block size. Measures of the size of the blocks were obtained by field counting; i.e., lists of seperate households were either made by the field personnel or obtained from directories.

At the block level, households were selected by full probability sampling methods. Each household had a known and equal chance of being selected. To give any adult a known probability of being interviewed within a household, a respondent-selection key was used that, with two eligibility questions, tells the interviewer whom to interview in each household. The American sample consists of 1,506 cases.

The German ALLBUS sample universe consists of the non-institutionalized population of the Federal Republic of Germany, including West Berlin. Selection of respondents was also by multi-level area probability sampling. At the first level, 630 electoral precints were randomly drawn from a list including all the electoral and synthetic precincts. The probability of selection was proportional to the number of private households in the precinct. At the second level, households were randomly selected from the electoral precincts. Finally, a single respondent was randomly selected from the elected household by means of a random number key employed by the interviewer. This selection procedure led to a representative sample of households since each individual household has approximately the same chance of being selected. Since the probability of a person living in a household being selected is inversely related to the

number of persons in that household, the German sample contain weights to adjust for sampling error. However, these weights are not used in the current study. The West German sample consists of 2,991 cases (Davis and Smith, 1983; Peterson, 1985).

Because of the collaboration between the two research centers, both surveys share a subset of common items (Peterson, 1985). The common items we analyze consist of seven questions about the respondent's approval or disapproval of legal abortion under different conditions. We reproduce both the English and German versions of these questions in Appendix A. Six items have been part of the NORC GSS since its inception in 1972. The conditions associated with the items are: (a) a strong chance of serious defect in the baby (DEFECT), (b) the woman is married and does not want any more children (NOMORE), (c) the woman's health is seriously endangered by the pregnancy (HEALTH), (d) the family has a very low income and cannot afford any more children (POOR), (e) the woman became pregnant as a result of rape (RAPE), and (f) the woman is not married and does not want to marry the man (SINGLE). The seventh item, introduced to the GSS in 1982, asks whether the respondent approves of legal abortion if the woman wants it for any reason (ANY). For the remainder of this study, we use the mnemonics to refer to these items. We limit our analysis to just those respondents who gave valid responses (Yes or No) to all seven items. This restriction reduces the

number of cases analyzed to 1,290 Americans and 2,068 West Germans.

6.2 Method

Recently, there have been two promising developments in the use of multiple items in the measurement of latent variables. The first is the application of latent class models to scale analysis and the second is the development of latent trait models to attitude data.

For the purposes of this study we first apply some latent trait extensions of the Guttman model to the seven abortion items. Starting with the simplest and most general non-parametric Mokken model, we will try to find unidimensional scales for each country, seperately. Once we find a unidimensional scale for the (sub)set of abortion items, we will perform a Mokken test to investigate the robustness of this scale across the two populations.

After the Mokken analysis, we will use various latent class scaling models which are presented as stochastic extensions of the Guttman scale model. A problem with these models is, however, that they are impractical to use for the analysis of more than four items. For this reason, it will become difficult to apply the latent class scaling models to the *seven* abortion items. To circumvent this problem, we will apply the latent class models to the particular set of items that meets the criteria for a robust non-parametric scale.

First, we will compare the Response Error Models and Intrinsically Unscalable Class versus Response Error Models without an unscalable class. This test will show the necessicity for including an unscalable class in the model.

Second, we will compare Proctor's model to the other Response Error Models. This test gives us information about the variety of error structures relative to Proctor's uniform error model for the abortion items in both countries.

Third, we will compare the "pure" Goodman model to hybrid latent class models that combine the response error type of models with an additional class of unscalable respondents. This test will provide us information about whether or not response errors are neccessary for the scale-type respondents.

Finally, we will apply the Multi-Group Latent Class model to the scaling model that fits best to the robust abortion items in both countries. This model gives us insight in the differences between the reliabilities of the indicators of the latent variable in each group.

Returning to one of the main questions of this research, the establishment of a unidimensional metric scale of the abortion items for Germany and the United States, we assign a metric to the ordinal Mokken scale. For this purpose, we use the log-multiplicative association methods.

Next, we will try to fit the one-parameter Rasch model. This parametric latent trait model is a specification of the

Mokken model and as such is used in the construction of a more refined scale of attitudes toward abortion. Since the conditional maximum likelihood estimation procedure is superior to unconditional maximum likelihood procedures, we will use the PML-computer program which provides this procedure of estimation.

Should the one-parameter model prove to be inadequate we will move on to the more complex two-parameter model. In the case of the two-parameter model we will use two competing and widely available computer programs: LOGIST and NOHARM. The comparability of goodness-of-fit of both models will be investigated.

Finally, we will develop a LISREL model that illustrates how the researcher can use the scale produced in the course of the scale analysis as a "reference indicator" in cross-population

6.3 Model Fit and Statistical Testing

In addition to the chi-square as statistical measure of goodness of fit, we will often use other measures of fit in comparing measurement models. As is well-known from statistics, a large sample size implies a increased power of the test so that small deviations from the model predictions cause already significant values of the test statistic, resulting in statistical rejection of the model. This is the classic situation of the choice between type I and type II errors in statistical testing. When α is chosen to be small,

the probability β of accepting the H_0 (the latent class or latent trait model) though it is false is gets larger and, as a consequence, the power of the test $(1-\beta)$ becomes smaller. Since both the American and West German sample sizes are relatively large (N=1290 and N=2068, respectively) in the present study we will often look at other criteria for evaluating a model (latent class as well as latent trait models). In particular, if a model does not fit the data statistically, we will look for additional criteria in order to evaluate the fit of the model. For example, inspection of the residuals gives information about the specific discrepancies between the observed and model-generated quantities (e.g., cell frequencies or covariances) or by checking the significance of the parameter values of each model we can detect which model provides the best solution for the research problem at hand. In Chapter 7, we will discuss the results of our comparative analyses for both hierarchical and non-hierarchical models.

6.4 Computer Programs Used for the Scale Analyses

As mentioned before, one of the advantages of latent trait models is that they can be falsified. We will show that, indeed, for most of the latent-trait models goodness-of-fit tests have been developed. These tests compare the observed data with the quantities predicted from the model. A widely used first-order indicator of misfit of the test *as a whole* is the chi-square statistic.

As argued earlier, we will not consider some arbritary significance level for rejection of a measurement model. Because we use relatively large samples in the present research, high values of the test statistic could be acceptable in circumstances where the test statistic does represent a large systematic departure from the model. For this reason we will also use exploratory techniques, if provided by the computer programs, to detect the sources of deviation of the measurement models.

In the next section, we will first discuss two estimation techniques commonly used for the estimation of the item and person parameters of the parametric latent trait models, the unconditional or joint maximum likelihood method and the conditional maximum likelihood method, respectively.

Subsequently, we will discuss the computer programs that are used for the assessment of the unidimensionality of the seven abortion items for the Americans and West Germans. The discussion will be focussed on the availability of goodness-of-fit tests and other diagnostic instruments offered by the different computer programs.

6.5 Estimation of Item and Person Parameters

The parameters of the latent trait model cannot be estimated without various assumptions such as unidimensionality and local independence have been made and the form of the ICC is specified based on the model that is

chosen. In the following sections we will discuss two estimation procedures; the Unconditional or Joint Maximum Likelihood Method and the Conditional Maximum Likelihood Method. We will demonstrate that the properties of these item parameter estimation procedures are inextricably intertwined with the computer programs used to implement them.

6.5.1 Joint Maximum Likelihood Estimation

In principle, the estimation of parameters by the method of maximum likelihood is straightforward. What is required by the maximum likelihood method is that the likelihood function be maximized.

However, the maximum likelihood estimation method fails for *simultaneous* estimation of item and person parameters in latent trait models. The first set of parameters are called the "structural" parameters" because they are common to sets of observation, while the person parameters are called "incidental parameters" because they increase with the number of observations (Neyman and Scott, 1948).

Neyman and Scott (1948), Andersen (1973), (see Hambleton and Swaminathan, 1985:127) point out that the presence of structural parameters affects the estimates of the person parameter. They showed that the maximum likelihood of the structural or item parameters do not converge to their true parameter values as the sample gets larger and larger; i.e., the estimates are not consistent.

This lack of consistent estimators is particularly problematic when the number of respondents increases because adding new respondents changes the number of parameters to be estimated. For instance, when N respondents take a test that has k items, the number of parameters to be estimated is 2k item parameters and N person parameters in the case of the two-parameter models, and 3k item parameters and N person parameters in the case of the three-parameter models. Hence, a total number of, respectively, 2k+N and 3k+N parameters must be estimated.

A problem arises at this point. Since the item parameters and person parameters are unobservable, there is a certain degree of indeterminacy in the model. In recent years this indeterminacy has been called the "identification" problem. In the two- and three-parameter logistic models this indeterminacy concerns the origin and unit of the scale of the parameter estimates whereas in the Rasch model there is only an indeterminacy in the origin of the scale.

Although many different schemes could be used the LOGIST computer program (Wingersky, Barton, and Lord, 1982) solves the identification problem by fixing the mean of the person parameter values at zero and the standard deviation to one. Consequently, in the two- and three-parameter models there are N+2k-2 and N+3k-2 free parameters to be estimated, respectively.

In addition to the statistical problems of simultaneously estimating the parameters, numerical problems arise that have to do with the fact that the ICC's of the logistic models are non-linear. This non-linearity results in the likelihood equations being nonlinear and solving a system of nonlinear equations needs very complex estimation procedures (Swaminathan, 1983). The numerical methods used may provide local minima instead of the absolute one (Samejama, 1973).

In the case of the one-parameter Rasch model, the estimation procedure is much simpler since the total number of items endorsed is a sufficient statistic for estimating the person parameter and the proportion of the respondents who endorsed the item is a sufficient statistic for estimating item difficulty. Therefore, by conditioning on the total number of items endorsed, the likelihood function for estimating the item parameters can be expressed in terms of the item parameters only, and consistency and unbiasedness is assured (Andersen, 1977). This so-called conditional maximum likelihood (CML) approach will be discussed in the next section.

6.5.2 Conditional Maximum Likelihood Estimation

Instead of estimating item and person parameters jointly, we can estimate the item parameters only. Andersen (1972, 1973) developed a maximum likelihood procedure that yields *consistent* estimates of the item parameters by

employing a conditional estimation procedure. Under this CML procedure, the likelihood function for estimating the item parameters is only expressed in terms of the item parameters by conditioning on the total scores. This procedure can be used only with the one-parameter Rasch model because only in this model is the total score a sufficient statistic for estimating the person parameter.

In contrast to the joint maximum likelihood procedure, the CML method provides estimates that are grounded on the principle of "specific objectivity"; i.e., the item parameters are estimated *separately* from the person parameters and vice versa. Under the model the total number of parameters to be estimated is $2(k-1)$. The model sets the unit of measurement at 1 ($b=1$) and the identification problem involves only the origin of the scale. Therefore, at the most $k-1$ item parameters have to be estimated under the Rasch model. In this study we will use the computer program PML (Gustaffson, 1979) to estimate the parameter values of the Rasch model for the seven abortion items. In PML the origin of the scale is determined by fixing the average item parameter value at zero. Regarding the person parameters, the model estimates $k-1$ person parameters for respondents who endorsed a total number of items between 1 and $k-1$.[18]

---

[18]Maximum likelihood estimation fails when a perfect score or a zero score is encountered because the maximum likelihood estimators for these cases are, plus or minus infinity, respectively. As a consequence, these cases are left out of the analysis. Bayesian estimation of item parameters (Swaminathan and Gifford, 1982) and Marginal Maximum Likelihood Estimation procedures (Thissen, 1982) are two methods that do not eliminate perfect response patterns.

Note that in the Rasch model any new item changes the number of parameters to be estimated. Note further, that adding new respondents does not change the number of parameters to be estimated, (assuming that each total score group is represented in the original sample), because the total number of items endorsed is a sufficient estimator of the person parameter. Only the CML procedure yields estimates of the item parameters (Andersen, 1973) (Fischer, 1974) that are known to be consistent.

In addition to consistent estimation, Gustaffson (1980) argues that there is a more important difference between the two approaches:

> "On the basis of the CML approach it is possible to devise efficient statistical tests of fit with known statistical properties, while under the UML approach only approximate statistical tests have been formulated". (Gustaffson, 1980:210).

Gustaffson (1980) made this procedure practical for sets with as many as 80 to 100 items. In addition, he developed the computer program PML (Gustaffson, 1979; Molenaar, 1981) which is suitable for CML estimation in the one-parameter model.

6.5.3 MOKKEN Program

In the present study, we begin with the construction of a stochastic Mokken scale'' of the abortion items for

---

''(cont'd) For a discussion of these two alternative approaches we refer to Hambleton and Swaminathan (1985).
''Standalone program written in FORTRAN 77 both for mainframe and micro computers is described by Kingma and Taerum (in press, a,b,).

Germany and the United States, separately. In this respect, the Mokken program provides a *non-parametric* procedure for testing the unidimensionality of this set of items. The Mokken method uses Loevinger's (1948) H index to assess the unidimensionality of the items both for the whole set of items and for each item, seperately (See section 5.2). Statistically based search procedures have been developed for adding or deleting items from a putative Mokken scale by maximization of the scale coefficient H.

Another feature of the Mokken methods, the Mokken test, allows the researcher to assess the goodness of fit of the Mokken model. By testing the null hypothesis that the pattern of relationships among items holds across two or more populations, we can test the robustness of the scale.'' Moreover, the Mokken test can be used to reduce a set of scalable items to a subset that is robust across the populations studied. For these reasons, the Mokken test will prove very useful to the present study since it will allow us to determine whether a scale or scales we construct are robust across the two countries studied.

The Mokken test provides various statistics for the whole scale a[illegible]he individual items. It involves the comparison of the values of H and H_i for different samples against an average or pooled value of H. Using the T-statistic developed by Mokken (1971:169), we can test the null hypothesis that the H and H_i coefficients are invariant

---

''Note, that the Mokken Test parallels the previously discussed Andersen test.

across the subpopulations studied. If we fail to reject the null hypothesis, we can infer that the Mokken model fits the data and use the average or pooled values of the indexes to produce a single value that applies to all groups. The T-statistic has a chi-square distribution with degrees of freedom equal to the number of groups minus one.

6.5.4 PML

As outlined above, only CML yields consistent estimates of the item parameters. In addition, the CML-method allows for the construction of goodness-of-fit tests which have at least asymptotically known distributions. The computer program PML (Gustafsson, 1979; Molenaar, 1981) employs the CML-algorithm for estimating the item parameters of the Rasch model and has incorporated different goodness-of-fit tests. Most of the tests concentrate on the property of "specific objectivity" and are developed to detect violations from the four basic axioms of the Rasch model (see Fischer, 1974:194):

1. Monotonicity; the probability of a positive response on an item is a strictly monotone function in the latent trait θ;
2. Dimensionality: the items measure the same latent trait;
3. Local independence: given the respondents' position on the latent variable, the probabilities of positively answering the items are independent;
4. Sufficiency of the total score: the total number of

items endorsed is sufficient for estimating the person parameter.

The best-known statistical test of the Rasch model is the conditional likelihood ratio test of model fit introduced by Andersen (1973). In this test procedure, the total sample of persons is grouped according to total score. In the present research, we will partition both the American and the German sample according to the internal criterion high and low total score groups.

To compute the Andersen test, the item parameters are estimated within the total group of respondents, and also within at least two subgroups of respondents. The test is sensitive to violations of the assumptions of sufficiency and monotonicity. When the model holds, the item parameter estimates of every subsample must be the same within random fluctuations. If the model does not hold, alternative models that allow for differences in the slopes of the ICC's (for instance, the two and three-parameter models) can be investigated. Since the likelihood ratio test procedure computes item parameter estimates for each subgroup, a reasonably large sample is needed. This, however, is not a problem in the present study where the number of respondents for both the Americans and the Germans is reasonably high.

The Andersen test can also be applied to subgroups that are partitioned according to external criteria other than total score such as, sex, age or country. In such cases, the test is sensitive to violations of the assumption of

unidimensionality (Gustaffson, 1980), since violations would reflect the variation of item parameters across groups. According to the assumption of local independence, respondents who endorse the same total number of items will get the same θ value on the latent variable. If, however, the set of items measures more than one dimension, the items will not be independent and both the assumptions of local independence and unidimensionality are violated. In this study, will partition the American and German sample according to the criterion sex.''

Martin-Löf (1973) introduced another test statistic which is sensitive to heterogeneous slopes of the ICC's. As in the Andersen test the sample is divided into k-1 score groups. However, this test is computationally very different compared to the Andersen test (Gustaffson, 1979) The Martin-Löf test is used in this study because it gives us *global* information about the contribution of each score groups to the overall misfit of the Rasch model to the data.

Another test that will prove useful in the present study is a "test of homogeneity of two sets of items" (Martin-Löf, 1973; Gustaffson, 1979). This test provides a test of the hypothesis that two disjoint groups of items measure the same latent variable. To compute this test, the items must be ordered a priori according to a certain criterion. The seven abortion items will be divided

---

'' Van Den Wollenberg (1979:ch.3) and Gustaffson (1980) argue that when the correlations between the grouping criterion and the subsamples differ the detection of multidimensionality by the likelihood ratio test fails.

according to the following criteria: abortion for medical or hard reasons (RAPE, HEALTH, DEFECT) and abortion for social or soft reasons (NOMORE, POOR, SINGLE and ANY). This will give us insight in the validity of the commonly found two-dimensional presentation of the seven abortion items. The test is developed to detect violations of the assumption of unidimensionality because the person parameters must be invariant for any subset of items that meet the criteria of the Rasch Model.''

Finally, Fischer and Schleibechner (1970) introduced a pairwise T-test that provides information on the fit of an individual item in the scale. This test is sensitive to violations of the assumption of invariance of item discrimination parameters. The test compares the item parameters for the low and high raw score groups by computing a statistic T_i for each item. Although the T statistic suffers from some problems (see Van den Wollenberg, 1979:31-39), the test is useful as a heuristic device for detecting misfitting items.

The test statistics derived from the previously discussed goodness-of-fit tests all follow a chi-square distribution. As noted earlier, the use of only first-order indicators of misfit may prove to be problematic in the

---

''Van den Wollenberg (1982a and b) demonstrated that the test statistics discussed above, may be insensitive to violations of the unidimensionality axiom under certain circumstances. He developed two new test statistics for lack of invariance of item discrimination parameters and unidimensionality. Unfortunately, these statistics are not available in PML.

present because the chi-square test is almost certain to be significant with large samples. Consequently, the null hypothesis of local independence, and therefore the unidimensionality of the set of items, is almost certain to be rejected. Second, if any outliers are in the sample, the deviation of the observed values from those predicted by the the model are not normally distributed anymore. In addition, outliers can also affect the tests of misfit of the models.

Therefore, in order to judge the unidimensionality of the seven abortion items for the United States and Germany, we will also consider more exploratory data-analytic approaches, in particular, graphical methods (Gustaffson, 1980; Molenaar, 1983) and slope tests for each item within each score groups. In this respect, PML provides a graphical procedure developed by Gustaffson (1980) which plots the observed proportion of positive answers for each score group against the corresponding predicted proportion. For each item and for each score group, PML computes a binomial test of the difference between observed and predicted frequencies of positive answers, including Z-scores and one-sided tests combined across score groups (Molenaar, 1983).

In sum, the results of the Andersen test and Martin-Löf tests will be used as *general* measures for the (mis)fit of the Rasch model to the abortion data. We use additional tests first because the size of the American and German sample makes the tests very powerful, and second, these tests do not provide any information about the (mis)fit of a

specific item in the scale. These additional diagnostic instruments will be employed to get more *specific* information about the nature of the violations from the Rasch model of each of the seven abortion items.[20]

6.5.5 LOGIST

The LOGIST program will be employed in this research in case the Rasch Model fails to describe the abortion data. One important reason for the misfit of the Rasch model may be the variations of the slopes of the item tracelines. This hypothesis can be tested by Birnbaum's two-parameter model which allows for differences in the slopes of the item trace lines.

In North America, LOGIST is one of the most widely used computer programs for estimating the one-, two-, and three-parameter models. The program is based on the UML estimation procedure and does not provide any goodness-of-fit tests. The model parameters are estimated iteratively (given some starting value) until satisfactory convergence is obtained. For details of LOGIST we refer to the user manual (Wingersky, Barton and Lord, 1982).

6.5.6 NOHARM

NOHARM (*N*ormal *H*armonic *A*nalysis *R*obust *M*ethod) is an IBM PC computer program for fitting both unidimensional and

---

[20] Duncan (1984b), Clogg (1987) and Kelderman (1984) present the Rasch model as a log-linear model for an incomplete table. This model can be estimated by using SPSSX.

multidimensional normal ogive models and latent trait models to binary data (Fraser, 1980).[2] In this study, NOHARM will be used for a test of the two-parameter model estimates as provided by the computer program LOGIST.

The computer program NOHARM facilitates parameter estimates for the one- two- and three-parameter model founded on a conceptually appealing approach. As discussed in chapter 5, NOHARM is based on a method to fit latent trait models by the analysis of covariance structures. The program NOHARM computes estimates of parameters of the model by fitting the polynomial curve to the normal ogive, weighted by the normal density function. The loss function (least squares function) is minimized using a conjugate gradients minimization algorithm which continues iterating so long as the function value continues to decrease and the magnitude is larger than some small value which is set initially (McDonald, 1982; McDonald, 1967).

The program computes the residual covariances of the items, after fitting the model, and gives the root mean square of these as an overall measure of misfit of the model to the data (Fraser, 1980). The magnitudes of the residuals give an indication of how well the data is approximated by the unidimensional normal ogive model (McDonald, 1981). "If the root mean square residual is in the order of the typical standard error of the residuals (four times the reciprocal

---

[2]The name of the program refers to the robustness of the program against violations of the normality assumption of the latent trait (McDonald, 1982).

of the square root of the sample size) we have a rough indication that a refined test of significance would not reject the hypothesized model. An examination of the residual matrix for clusters of large residuals may indicate groups of items that do not fit the model, and suggest a means to modify it in order to improve the model" (Fraser, 1980:2). Second- or higher-order indicators of misfit (for instance, the residual covariances) may be one of the exploratory techniques that offer an alternative route to assessing the unidimensionality of the abortion items by means of the chi-square statistic (McDonald, 1981; Hattie, 1985; Van Den Wollenberg, 1982 a,b).

6.5.7 ANOASC

In Chapter 5 we considered a series of models that can be used to estimate metric scale scores for the ordinal Mokken scale values. We accomplish this analysis by using the computer program ANOASC (A Computer Program for the *An*alysis *o*f *A*ssociation in a Set of K I-by-J *C*onditional Tables) written by Shockey and Clogg (Shockey and Clogg, 1983). The program ANOASC is based on an extension of Goodman's algorithm (Goodman, 1979b; Clogg, 1982) of multiway cross-classifications having ordered categories. All the log-multiplicative association models discussed in chapter 5 can be fit using ANOASC.

6.5.8 MLLSA

In the present study the computer program MLLSA (Maximum Likelihood Latent Structure Analysis) (Clogg, 1977), will be used to test the latent class scaling models discussed in Chapter 4. The restrictions that can presently be considered with MLLSA are of the following types:

1. Equality restrictions on the conditional probabilities; for instance, $\pi^{\bar{a}}_{1,x_1} = \pi^{\bar{b}}_{1,x_1}$.
2. Restrictions of parameters to fixed constants; $\pi^{\bar{a}}_{1,x_1} = 1$.
3. Equality restrictions on the latent class proportions; $\pi_1 = \pi_2$.
4. Restrictions that form a combination of type 1-3.

A maximum likelihood procedure for estimating the parameters of the general latent class model is described by Goodman (1974a) and Clogg (1977) has implemented these methods in MLLSA. The latest version of MLLSA also provides a procedure for determining the identifiability of the parameter estimates. This problem of identifiability refers to a dependency among the parameter estimates which means that more than one set of parameters can account for the data. This indeterminacy gives rise to difficulties in the interpretation of the model parameters and the program may take considerable time to converge. The latest version of MLLSA enables the researcher to test latent structure models *across* groups.

7. RESULTS OF THE "NON-PARAMETRIC ANALYSES"

7.1 Results of the Mokken Analyses

The proportion of respondents who endorse abortion under the seven conditions are given for Americans and West Germans in Table 7.1. Three features of these results warrant our attention: First, the items fall into two groups of low and medium difficulty for both Americans and West Germans. Following Mooijaart (1982), we refer to the low difficulty items as abortion for "medical" reasons (HEALTH, DEFECT, and RAPE) and the medium difficulty items as abortion for "social" reasons (POOR, SINGLE, NOMORE, and ANY). The difficulty of the medical items ranges from .85 (DEFECT) to .91 (HEALTH) for Americans and from .91 (RAPE) to .94 (HEALTH) for West Germans; the difficulty of the social items ranges from .44 (ANY) to .54 (POOR) for Americans and from .32 (ANY) to .53 (POOR) for West Germans. The question these results raise is whether the social and medical items represent one or two dimensions.

Second, although Americans and West Germans generally produce the same item orderings, there are two reversals: one for RAPE and DEFECT; the other for SINGLE and NOMORE. Whereas the DEFECT item is slightly more difficult than RAPE for Americans (.85 versus .87), it is slightly less difficult for West Germans (.93 versus .91). Whereas the NOMORE item is slightly more difficult than SINGLE for Americans (.50 versus .51), it is much less difficult for

Table 7.1: The Distribution of the Abortion Items for Americans and West Germans

Item	Difficulty Level*	
	Americans	Germans
ANY	.44	.32
SINGLE	.51	.34
NOMORE	.50	.45
POOR	.54	.53
DEFECT	.85	.93
RAPE	.87	.91
HEALTH	.91	.94
	N=1,290	N=2,068

*Proportion of respondents who endorse the abortion item.

West Germans (.45 versus .34). These reversals raise the question of whether the scale (or scales) that emerge from separate Mokken analyses of these items for the American and West German samples will be robust across the two populations.

Third, when we compare American and West German attitudes toward abortion, it appears that Americans hold a more liberal view with respect to the social items but a slightly more conservative view with respect to the medical items. The problem with interpreting these results, however, is that comparisons of the item marginals confounds differences between Americans' and West Germans' general attitudes, with the cross-national differences between the difficulty of *specific* items. If we can find a *scale* (or scales) that is robust for the two populations, we can use it to make a more straightforward comparison of abortion attitudes.

When we apply the Mokken method separately to American and West German items, we find the following: First, the P and P_0 matrices support the assumption of double monotony. These matrices are given in Tables 1.B and 2.B in Appendix B. Since none of the comparisons among the cell entries in the matrices yield differences opposite to the ones predicted by the Mokken model and the item difficulties, we regard these results as an indication that the seven items measure a single dimension in both countries.

The results for Loevinger's (1948) index of homogeneity are presented in Table 7.2. The coefficients H_i for all seven items are all quite high. They range from .82 for DEFECT to .89 for ANY for the American sample, and from .73 for RAPE to .79 for POOR for the West Germans. As a consequence, the H-coefficient of homogeneity for the entire set of items is .846 for the American sample and .773 for the West German sample. When we test these coefficients statistically, we find, first, that all are significantly greater than zero. Thus, we can reject the (weak) null hypothesis that the seven abortion items are independent. In addition, we also find that the coefficients are significantly greater than .60, the stringent cutoff value that Mokken (1971:185) uses to distinguish "strong" from "medium" Mokken scales. Thus, these results answer the first question raised above by indicating that the seven abortion items form a unidimensional scale in both countries.

We complete our scale analysis by looking at the results of the Mokken test we used to see whether the seven items form a scale that is robust across the two

populations. Table 7.3 contains the T-statistic, which has an approximate chi-square distribution with one degree of freedom for seven, six, five and four-item scales along with the item deleted from the scale.

Table 7.2: The H_i Values of the Abortion Items for Americans and West-Germans

	H_i	
	Americans	West Germans
ANY	.89	.78
SINGLE	.84	.78
NOMORE	.84	.78
POOR	.83	.79
DEFECT	.82	.75
RAPE	.84	.73
HEALTH	.87	.74

Table 7.3: Impact of Dropping Items on the T-Statistic for the Mokken Test

Items Dropped From the Scale	T-statistic of Robustness
None	20.20
ANY	9.89
ANY, HEALTH	6.03
ANY, HEALTH, RAPE	1.34

After trying different possible combinations of scales, we settled on the four-item scale since the value of the T-statistic is 1.34 which, for one degree of freedom, represents an excellent degree of fit. In contrast, as Table 7.3 shows, neither the five-, six-, or seven-item scales are as robust. Thus, although all seven items scale when both samples are analyzed separately, the results of our Mokken test show that a scale which is equivalent or robust across the two populations is a four-item scale which contains the items DEFECT, POOR, NOMORE, and SINGLE.

Since the four-item scale is robust across the American and West German populations, we can obtain better estimates of the scale's characteristics by combining the American and West German samples into a single group. When we do this, we find that the item difficulties are .41, .47, .53, and .90 for SINGLE, NOMORE, POOR, and DEFECT, respectively. The homogeneity coefficients H_i for these items are .80, .78, .79, and .86, and the coefficient for the scale as a whole H is .80.

7.2 Results of the Latent Class Analyses

In this study, the applications of a variety of latent class models should be regarded as a complementary analysis to the latent trait analyses. In this respect, we defined four tests:

1. Proctor's uniform error model versus response error models which have a variety of error structures;
2. Response error models and intrinsically unscalable class versus response error models without an unscalable

class;

3. The "pure" Goodman model versus the Dayton and Macready type models; and
4. Whether the latent distribution of the latent variable is the same for both countries. In addition, we will test whether the error rates are the same for both countries.

The chi-squared statistics of the latent class scaling models for Americans and West Germans are given in the Table 7.4. Note that the fit for the models *without* an unscalable class (model M_1, M_3, M_5, M_7 and M_9) is quite poor for both the American and West German sample and that the latent distributions are pretty much the same, regardless of which response error model is chosen. Moreover, the percentage of respondents in the scale types II and III is very small (see Table C.1, in Appendix C).

These models are all generalizations of Proctor's model and allow for a variety of error-structures. M_3 is the item-specific-error rate model which allows the conditional probabilities to vary from item to item within a scale type. M_3 applied to the four robust abortion items yield a L^2 of 27.6 on 7 degrees of freedom for the Americans. In order to answer the first question raised above, we compare the fit of Proctor's model to the fit of the response error model M_3, M_5, M_7, and M_9 for both countries. It is easy to see that the differences in chi-square statistics are highly significant for the Americans and the West Germans.

Table 7.4: Chi-Square Values for Latent Class Models Applied to Four Robust Abortion Items for Americans and West-Germans

		UNITED STATES	WEST-GERMANY
Model	DF	Likelihood-Ratio Chi Square	Likelihood-Ratio Chi-square
M_1	10	53.6	119.0
M_2	6	10.9	11.8
M_3	7	27.6	30.6
M_4	3	3.0	26.0
M_5	7	24.5	95.8
M_6	3	2.2	11.9
M_7	9	42.0	107.3
M_8	4	10.7	11.8
M_9	5	26.5	27.9
M_{10}	1	3.0	5.4
M_{11}	7	15.8	11.8

Explanation of models:

M_1 Proctor's Model
M_2 Proctor's Model and Unscalable Class (UC)
M_3 Item-Specific Error Rate Model
M_4 Item-Specific Error Rate Model and UC
M_5 Type-Specific Error Rate Model
M_6 Type-Specific Error Rate Model and UC
M_7 False-Positive/False-Negative Error Model
M_8 False-Positive/False-Negative Error Model and UC
M_9 Latent Distance Model
M_{10} Latent-Distance Model and UC
M_{11} Goodman's model

For instance, the type-specific error rate model yields the greatest improvement in fit compared to the fit of Proctor's model (53.6-24.5=29.1 which with 10-7=3 degrees of freedom is highly significant) for the Americans. For the West Germans the latent distance model M_9 appears to result in the greatest improvement of fit (119.0-27.9=91.1 with 10-5=5 degrees of freedom which is highly significant).

Thus, these results indicate that despite the fact that none of the response error models have fit the data satisfactorily, the fit of the restrictive Proctor model can be improved by allowing the error rates to vary.

Turning to the second test, we try to improve the fit of the models by relaxing the assumption of population homogeneity. By assuming the existence of an intrinsically unscalable class whose members are viewed as responding randomly or else in terms of item orderings that differ from the dominant one, we test the response error models plus unscalable class versus response error models only. Table 7.4 present the extended models for the United States and West Germany, respectively. It is obvious that the extended models M_2, M_4, M_6, M_8, and M_{10} for the Americans, and the models M_2, M_6, M_8, and M_{10} for the Germans provide a significantly better fit to the data relative to the same models but without an unscalable class (see Table 7.4).

Further, when we look at the corresponding L^2 values for the models M_2, M_4, M_6, and M_{10} in Table 7.4, we see that these models describe the data very adequately for the American sample ($p>.05$). Since we prefer the type of model which strikes a balance between goodness-of-fit and parsimony, we select M_2 as being the best model for representing the abortion data. This simple model provides an excellent fit ($L^2=10.9$) and in terms of parsimony, M_2 has more degrees of freedom than the other fitting models M_4, M_6, and M_{10}. Moreover, under this model the estimated proportion of cases classified as unscalable is relatively low. We find that 16 percent of the Americans can be

regarded as having response patterns that are intrinsically unscalable and that therefore the unidimensionality of the four abortion items holds for a reasonable part of the American population. Further, when we look at Table C.1., we see that the proportion estimated under first scale type (1,1,1,1) is 43 percent for the Americans which is relatively high and may be indicative for "ideological" approval of abortion (Duncan, 1982).[23] Finally, the response error rate *a* for each item is estimated at .01 which means that, given latent class, the probability of response error is very low.

Next we consider the German sample in Table 7.4. The results show that that only model M_2 and $M_{1,1}$ fit the the abortion items according to the conventional statistical criterion of $p>.05$ ($L^2=11.8$ with 6 degrees of freedom and $L^2=11.8$ with 7 degrees of freedom, respectively). Thus both models describe the data adequately; however, $M_{1,1}$ is more parsimonious in its number of parameters. Further, the estimated error rate under Model M_2 is 0.0. This indicates that Proctor's model and unscalable class is reduced to Goodman's model $M_{1,1}$ where no response error is expected to occur in the responses of the scalable class.

The estimated proportion for the unscalable part of the populations is given in the last colomn of Table C.2. We see

---

[23] "Ideological" respondents are persons who consistently endorse or reject any item that pertain to the attitude being measured. For instance, ideological respondents would endorse or reject legal abortion, regardless of the grounds offered for its justification. In another study, we try to seperate these respondents from those with less extreme positions.

immediately that for both models M_2 and M_{11}, 30 percent of the population belong to the unscalable class and that this percentage does not differ very much for the hybrid latent class model M_4, M_6, and M_8. These results mean that the responses to the four abortion items are independent for a substantial part of the population and that this proportion does not drop when we relax the assumption that members of the scalable classes response according Guttman's perfect scaling model.[23] Further, when we look at the percentage of respondents who respond positively to all four abortion items in Table C.1 (scale type I), we see that these percentages are much lower than for the Americans. Therefore, it is very likely that the number of "ideological" respondents is much higher in the American sample relative to that of the West Germans.

These results answer the second question raised above by indicating that for both the Americans and the West Germans the fit of the models has been improved by including an intrinsically unscalable class in the models.

Turning to the third question, Table 7.4 shows that for the American sample, Goodman's model (M_{11}) does not fit the data in a reasonable way ($L^2=15.8$ with 7 degrees of freedom which is significant $p<.05$). Since model M_{11} is not demonstrably superior to M_2 we conclude that Proctor's model and unscalable class represents the abortion data

---

[23]Since we already suspect that the four items do form a unidimensional scale for the West Germans, these results may indicate that the error structures, as defined in the latent class models, are not flexible enough. That is, the program does not allow us to specify unequality constraints on the conditional probabilities.

adequately.

In the case of the West Germans, both model M_2 and $M_{1,1}$ fit the abortion data very well. However, since $M_{1,1}$ is most parsimonious and because the error rate is 0.0 under M_2, the "pure" Goodman model is an appropiate scaling model for the four abortion items.

Answering the first part of question four, we first have to decide on a scale model which is valid for both countries. We have shown above that Proctor's model represents a very acceptable fit for both the American and West German samples. (Note that for the Germans, we treat Goodman's model as a Proctor model plus an unscalable class with an error rate of 0.0.). We imposed homogeneity constraints on the latent class proportions, that is, we assumed that the error rates do not depend on country. This restricted-homogenous model does not fit the data (L^2=27.5 on 11 degrees of freedom) which signifies that the distribution of the latent variable is not constant across the United States and West Germany.

7.3 Results of the ANOASC Analyses

As we have seen, the results of the Mokken Test indicate that the items DEFECT, POOR, NOMORE, and SINGLE form a four-item scale of attitudes toward abortion that is robust across the American and West German population. Latent class analysis provided some convergent evidence of the unidimensionality of the four abortion items: the robust scale items fit the Guttman-type models well although the proportion of respondents in the unscalable class is very

large for the German sample. In the following, we will refine the four item robust scale by developing metric scale values.

Discussing the log-multiplicative association model in Chapter 4, we mentioned the importance of an "instrumental" variable for the estimation of the categories of the scale. In the present research, the variable church attendence (CHATTEND) is used as criterion variable, partially because church attendence is the only variable that correlates with both American and West German attitudes toward abortion.

In both countries, church attendence was measured by "frequency of church attendence" by a scale ranging from never (0) to more than once a week (8) for the Americans, and from more than once a week (1) to never (6) for the Germans.

Americans

0. never
1. < 1 year
2. 2 times a year
3. several times a year
4. once a month
5. 2-3 times a month
6. nearly every week
7. every week
8. more than once a week

West Germans

1. more than once a week
2. once a week
3. 1-3 times a month
4. few times a year
5. seldom
6. never

We recoded the American and West German classification in the following way:

Recode	Americans	West Germans
1. never	0=1	6=1
2. seldom	1,2,3=2	4,5=2
3. often	4,5=3	3=3
4. very often	6,7,8=4	1,2=4

The model of conditional independence (attitude toward abortion independent of CHATTEND for each country) yields L^2=642.88 with 24 degrees of freedom, which is a very significant result. Since this model does not fit the data, we can use the variable CHATTEND as an instrument to calibrate the variable "attitudes toward abortion'.

The heterogeneous column, row-column effects model (church attendence-attitudes toward abortion association depends on the frequency of CHATTEND, but the attitude toward abortion scale scores are not allowed to differ across the Americans and the West Germans) presents a L^2=14.3 with 15 degrees of freedom. The goodness of fit of the latter model compared to the goodness of fit of the heterogeneous row-column, row-column effect (L^2=11.9 with 12 degrees of freedom) gives us an idea of the homogeneity of the scale scores of the "attitude toward abortion" scale across the two countries. Since this latter model does not provide a significant improvement in fit compared to the heterogeneous column, row-column effects model (L^2=14.3-11.9=2.4 with 15-12=3 degrees of freedom), there is no reason to reject the model with homogeneous abortion scale scores across the countries. Thus, since the estimates of the parameters for scores on the "attitudes toward abortion scale" μ_i are constrained to be equal for the Americans and West Germans, the estimated differences between the scale scores 0, 1, 2, 3, and 4 are also assumed to be equal under the model. The estimated category scores $\hat{\mu}_i$ for the metric scale are, -.75, -.17, .07, .26, and .58, respectively, whereas the distances between the intervals

are:

$\hat{\mu}_1 - \hat{\mu}_0 = .58$

$\hat{\mu}_2 - \hat{\mu}_1 = .23$

$\hat{\mu}_3 - \hat{\mu}_2 = .20$

$\hat{\mu}_4 - \hat{\mu}_3 = .32$

The estimated difference between the scale categories 0 and 1 is about two and a half times the distance between categories 2 and 1 (.58/.23=2.52), the distance between the categories 2 and 1 is about the same as that between the categories 3 and 2 (.23/.20=1.15), and the distance between the categories 4 and 3 is about one and half the distance between the categories 3 and 2 (.32/.20=1.6). Thus, the differences between the extreme items are much larger compared to those in the middle categories. This is exactly what we should expect if a unidimensional scale is well constructed; i.e., there are only a few respondents who will answer the easiest item (DEFECT) negatively and the hardest item (SINGLE) positively (Taylor, 1983).

Finally, the hypothesis of equal intervals between the scores of the "attitude toward abortion" scale gives $L^2=39.15$ with 18 degrees of freedom. The difference between $L^2 39.15 - L^2 14.30 = L^2 24.85$ on 3 degrees of freedom which statistically significant. This result indicates that the ratio's of distances, as presented above, are significantly different from one. In sum, the results of the ANOASC analyses indicate, that we can refine our equivalent Mokken scale by assigning any scores to the ordinal scale scores that preserve the ratio of distances as estimated under the heterogeneous column, row-column effects model.

7.4 American and West-German attitudes toward abortion

Referring to the third question raised at the beginning of this chapter, we now can compare the American and West-German attitudes on the refined metric four-item scale. The mean of the recalibrated scale scores for Americans is .13, while the mean for West-Germans is .10. Thus, Americans are slightly more likely to approve of legal abortion. However, this difference is not statistically significant ($p>.05$). In the case of the F-test, we find that the variances of the abortion variable significantly differ for the Americans and West-Germans ($p<.01$). These variances are .22 and .14 for the Americans and West-Germans, respectively.

These findings can be explained by the results presented in the Tables 7.1 and 7.5. Table 7.1 shows, that proportionally more Americans hold a liberal view with respect to the social items but proportionally more West-Germans hold a more conservative view with respect to the medical items. As noted before, many researchers consider these differences in item marginals as evidence for the bi-dimensional solution of the abortion items: attitudes toward the use of abortion for medical reasons and attitudes toward the use of abortion for social reasons. However, the results of the comparison of the *scale* scores suggest that this conclusion is an artifact of the confounding differences between the Americans' and West-Germans' positions on the latent variable with the cross-national differences between the item's positions on the latent variable.

Table 7.5: Distribution of the Non-Metric and Metric Scale scores for the Americans and West-Germans.

Non-Metric Scale Scores:		0	1	2	3	4
Metric Scale Scores:		-.75	-.17	.07	.26	.58
		AMERICANS (N=1290)				
		172	342	111	112	553
	%	13.3	26.5	8.6	8.7	42.9
		WEST-GERMANS (N=2068)				
		138	701	335	308	586
	%	6.7	33.9	16.2	14.9	28.3

Table 7.5 shows the distribution of the non-metric and metric scale scores for the Americans and West Germans of the four abortion items, DEFECT, POOR, NOMORE, and SINGLE. Note the differences in variation between the two populations; i.e., the percentage of cases in *both* extreme categories is greater for the Americans than for the West-Germans.[24]

Furthermore, since the distances between the extreme scale scores are larger than the distances of the intervals between the intermediate scale scores, and given the distribution of the scale for both populations, the

---

[24] The F-max test is used to test the hypothesis of homogeneity of variance (Winer, 1970:99). The result indicates that the value of the F-max statistic is not significant ($p>.10$).

differences between the means are not significant anymore. Thus, the difference between the variability of the scores and the fact that we use recalibrated scales explains the results obtained by the T- and F-test.

Of course, it can be argued that our metric scale scores depend on the particular instrumental variable CHATTEND chosen. Therefore, we validated the results of the ANOASC scale analyses by the use of different instrumental variables. Unfortunately, suitable instrument variables other than CHATTEND were only available in the American NORC Survey. We considered the following instrumental variables for the Americans:

1. What do you think is the ideal number of children for a family to have? (CHLDIDEL)
2. There has been a lot of discussion about the way morals and attitudes about sex are changing in this country. If a man and woman have sex relations before marriage, do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all? (PREMARSEX).
3. Do you expect to have any (more) children? How many more? (CHLDNUM).
4. What is your opinion about a married person having sexual relations with someone other than the marriage partner- - is it always wrong, almost wrong, wrong only sometimes, or not wrong at all? (XMARSEX)
5. What about sexual relations between two adults of the same sex-- do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all? (HOMOSEX).

Table 7.6 presents the metric scale scores of the five ordinal abortion scale values for the American sample. The model to be tested was that the association between the variable "attitude toward abortion" and the instrumental variable(s) exhibits linear-by-linear interaction. Note that this model is the same as model (13) in Chapter 5 but without the group variable "country".

Table 7.6 shows that the model provides an excellent fit for the five instrumental variables, CHLDIDEL, PREMARSEX, CHLDNUM, XMARSEX, and HOMOSEX.

Table 7.6: Estimated Scale Scores for the Abortion Attitude by Attitude on CHLDIDEL, CHLDNUM, PREMARSEX, XMARSEX, and HOMOSEX

CHLDIDEL

Abortion Attitude Scale Scores	Metric Scale Scores $\hat{\mu}_i$	Differences $\hat{\mu}_{i+1}-\hat{\mu}_i$
0	-.68	.44
1	-.24	.44
2	.20	-.13
3	.07	.59
4	.66	

L^2= 21.9 on 12 df.

PREMARSEX

Abortion Attitude Scale Scores	Metric Scale Scores $\hat{\mu}_i$	Differences $\hat{\mu}_{i+1}-\hat{\mu}_i$
0	-.69	.42
1	-.27	.50
2	.23	-.11
3	.12	.50
4	.62	

L^2=5.7 on 6 d.f.

CHLDNUM

Abortion Attitude Scale Scores	Metric Scale Scores $\hat{\mu}_i$	Differences $\hat{\mu}_{i+1}-\hat{\mu}_i$
0	-.69	.35
1	-.34	.35
2	.01	.27
3	.28	.38
4	.66	

L^2=6.31 on 9 d.f.

Table continued

XMARSEX

Abortion Attitude Scale Scores	Metric Scale Scores $\hat{\mu}_i$	Differences $\hat{\mu}_{i+1}-\hat{\mu}_i$
0	-.62	.30
1	-.32	.36
2	.04	.18
3	.22	.46
4	.68	

L^2=7.03 on 6 df.

HOMOSEX

Abortion Attitude Scale Scores	Metric Scale Scores $\hat{\mu}_i$	Differences $\hat{\mu}_{i+1}-\hat{\mu}_i$
0	-.69	.49
1	-.20	.24
2	.04	.18
3	.22	.44
4	.66	

L^2=5.26 on 6 df.

Note that the instrumental variables CHLDIDEL and PREMARSEX have reversals in the scale categories 3 and 2, (-.13, and -.11, respectively) and, as a consequence, may not be considered as suitable instrumental variables. These reversals, however, may be attributed to correlated measurement error because CHILDIDEL and PREMARSEX both relate to specific items in the scale: NOMORE, and SINGLE, respectively.

The results of the recalibration of the abortion scale scores by CHLDNUM, XMARSEX, and HOMOSEX, show that for XMARSEX the distance between the scale scores 2 and 1 is larger than the distance between the scale scores 0 and 1, whereas for CHATTEND, the reverse is true. A more important finding is, however, that the extreme difference between the

scale scores 0 and 1 no longer occurs. Therefore, we want to examine if the differences in attitudes toward abortion between the Americans and West-Germans are an artifact of the particular calibration CHATTEND.

First, we assume that the intervals are the same for both Americans and West-Germans; the equal interval specification (uncalibrated scores). Subsequently, we assume that the intervals between the categories of the the recalibrated abortion scores by the instrumental variables CHLDNUM, XMARSEX, and HOMOSEX are the same for both countries.[28] Table 7.7 presents the results for the T and F-test of the comparisons with respect to the choice of different instrumental variables.

Table 7.7: Results of the T and F-Test for the Instrumental Variables none, CHLDNUM, XMARSEX, and HOMOSEX.

Instrumental Variable	Difference Between Means (T-test)	Difference Between Variances (F-Test)
None	1.33*	3.21*
CLDNUM	1.35*	3.36*
XMARSEX	3.89*	1.35*
HOMOSEX	3.08*	3.08*

* $p < 0.00$

For all calibrations both the means and the variances significantly differ for both countries. In all cases, we find that the means as well as the variances are somewhat

---

[28] Unfortunately, other instrumental variables than CHATTEND were not available for the West-Germans. Therefore, we use the American calibrations for the West-Germans.

higher for the Americans. When we compare these findings to the results of the comparison discussed in Section 7.3, we may conclude that the variation in attitude toward abortion is greater for the Americans than for the West-Germans. In addition, this result does not depend on the particular instrumental variable chosen. Except for the calibration by CHATTEND, we find evidence for the difference between the means of the scale scores; i.e., Americans are slightly more liberal in their attitudes toward abortion.

However, the estimates of the distances between the calibrated scale scores by CHATTEND appear not to be robust; i.e., the extreme difference between the categories 0 and 1 for the robust metric scale with CHATTEND as instrumental variable, does not exist for the calibrations of the abortion scale scores by non-metric scores, CHLNUM, XMARSEX, or HOMOSEX. As a consequence, the means do not differ significantly for the recalibrated scale by CHATTEND, whereas for the other recalibrations the means differ significantly for both populations.

We conclude by emphasizing that only a comparison of *scale* scores suggest that both the variances and the means of the abortion scale scores differ for the Americans and West-Germans. Looking at individual items do not provide that information since it confounds the difference in the general attitude toward abortion measured by a set of items with the difference in the difficulty of specific items. For this reason, the two-dimensional solution of the abortion items is an artifact of the differences in variations in abortion attitudes for Americans and West-Germans.

7.5 Conclusion

Using the Mokken method and Mokken Test, we have developed a four-item scale of attitudes toward abortion that is robust across the American and West-German populations.

Latent class scaling models are analyzed for the four robust items. The Proctor model augmented with an unscalable class provided the best fit for the Americans, whereas the Goodman model was optimal for the West-German sample. Multi-group latent class analyses showed that the error rates of the items depend on country.

Subsequently, log-multiplicative association methods were used to assign a metric to the categories of the robust Mokken scale. Comparing *scale* results of different recalibrations, we concluded that; Americans are more liberal in their attitudes towards abortion than West-Germans; Variation in attitude is greater for Americans than for West-Germans. Validation of the recalibrated abortion scale scores by CHATTEND, revealed that the estimates of the distances between the scale scores were not robust against the choice of the other instrumental variables: CHLDNUM, XMARSEX, and HOMOSEX, respectively.

8. RESULTS OF THE PARAMETRIC ANALYSES

Before discussing the results of the Rasch analysis, we will first deal with one of the limitations of maximum likelihood estimation procedures which was encountered in the dataset we use for this study. Maximum likelihood estimators are only consistent, efficient, and sufficient if they are finite. However, the person parameters for respondents who endorse all or none of the seven abortion items is positive or negative infinite, respectively. To cope with this problem, most of the computer programs which estimate the parameters of the one- and two parameter models, remove those respondents from the dataset before estimating the model parameters (see footnote 14, Chapter 6).

Clearly, this is not desirable in practical situations since every respondent should receive a finite score so that she can be included in the statistical analyses of these scores. In the present research, this artifact of the maximum likelihood estimation procedure means that for the Americans at least 88+500=588 respondents are left out of the analyses whereas for the Germans this number is at least 73+522=595 (see Tables C.3 and C.4), respectively). It is needless to say, that the effect of deleting such a large number of respondents may obscure the nature of the results.

In particular, in attitude measurement zero and perfect responses will often occur for at least two reasons: first, the number of items in attitude survey research is often small which simply increases the chance on a respondent's approval or disapproval of all the questions being asked and second, in attitude measurement it will often occur that

respondents either endorse or consistently reject all attitude items due to their "ideological" positions on the attitude being measured.[24] While it may be justified to exclude the "ideological" respondents from the analyses because they do not provide information on the degree of conviction,[27] it is a doubtful practice to delete *every* respondent with a zero or perfect response pattern from the scaling analysis.

Turning to the discussion of the Rasch results of our analyses, we will focus on the the various goodness-of-fit tests that are currently available in the computer program PML. In Chapter 6 we noticed that most of these tests are based on a partitioning of the dataset according to the criterion "number of items endorsed". We argued there, that if the principle of "specific objectivity" holds, the item parameter estimates of the subsamples and the total sample must be equal within chance limits. In the following we will discuss the results of the Rasch analyses of the seven abortion items for the Americans and West Germans based on the following goodness-of-tests:

1. Andersen and Martin-Löf tests to get a global idea of the fit of the Rasch model to the abortion items;

---

[24] These high percentages of zero and perfect responses are not typical for the abortion items. For instance, 30 % of the responses to the "civil liberty items" in the GSS 1972-1984 survey were zero or perfect responses; in the Quality of American Life (1978) Survey, 26 % of the respondents endorsed all or none of the "self-feeling" items, and Lucke and Schuessler (1987) found that 39 % of their German sample agreed or disagreed with all five "social life feelings" items.

[27] The omission of corner cells is consistent with the idea of the quasi-independence model: hence the similarity of Rasch scaling and log-linear analysis (Duncan, 1977).

2. The binomial test BINO per item scoregroup to detect specific items that violate the assumptions of the Rasch model.
3. Molenaar's U_k test statistic which circumvents the problem of differences in power of the score groups.
4. The Fischer-Schleibecher test is used as complementary instrument to the binomial test BINO in order to single out items that badly fit the Rasch model.

Table 8.1 presents the results of the Andersen and Martin-Löf tests of the seven abortion items for the American sample (X^2=32.09 on 6 degrees of freedom). The sample is divided into a low (sum scores 1 to 4) and a high (sum score 5 to 6) scoregroup, in such a way, that an asymptotically chi-square distributed test statistic can be formed.

Table 8.1: Results from the Martin-Löf Test and Andersen Test of the Seven Abortion Items for the Americans.

Martin-Löf-Test		
Score Group	Number of Observations	Contribution
1	40	1.94
2	94	5.14
3	292	5.11
4	92	8.18
5	97	18.04
6	87	119.54
T=157.95, df=30, p=.00		

Results from Andersen Test of equality of the item-parameters for low and high score groups:
X^2=32.09, df=6, p=.00

Table 8.1 also provides the results of the Martin-Löf test. The results indicate that score group 6 (containing those respondents who endorse six abortion items) contribute

the most to the total chi-square sum of T=157.94 with 30 degrees of freedom which is highly significant. The results of the two tests indicate that the item parameters are not invariant for the subgroups as defined by the tests and that we need to look for specific items that violate the assumptions of the Rasch model. For the Germans, Table 8.2 provides the results of both tests.

Table 8.2: Results from the Martin-Löf Test and Andersen Test of the Seven Abortion Items for West-Germans

Martin-Löf-Test		
Score Group	Number of Observations	Contribution
1	45	3.55
2	122	16.72
3	587	10.67
4	316	24.18
5	257	39.10
6	146	6.25
T=100.48, df=30, p=.00		
Results from Andersen Test of equality of the item-parameters for males and females: X^2=30.73, df=6, p=.54		

The value of 30.73 for χ^2 with 6 degrees of freedom for the Andersen Test, and the T-value of 100.48 with 30 degrees of freedom for the Martin Löf test is in both cases highly significant. Thus the results of these two global goodnessof-fit tests indicate for both countries that the item parameters are not invariant for the subgroups as defined by the tests. Even if we take the relative large samples of the Americans and (in particular) the Germans into account, the test results are not satisfactory.

As mentioned in chapter 6, the computer program PML provides a graphical representation for each items of the observed proportion of positive responses for each score group, against the corresponding predicted proportion. If the items fit the Rasch model, the points in the scatterplot are evenly distributed around the main diagonal of the graph. If the observed proportion of respondents is larger than the predicted proportion for the lower score groups is smaller than the predicted proportion for the higher score groups, the slope of the item trace line may be relatively steep; i.e., the discrimination power of the item is too high relative to the other items. If the opposite pattern is observed, the slope of the item trace could be relatively flat and, therefore, the discrimination power of the item is low relative to the other items (Molenaar, 1982). PML provides (for each score group) and each item a binomial test of the difference between the observed and the predicted frequency.

Table C.5 presents the results of the binomial test BINO of the seven abortion items for the Americans. The results of the one-sided binomial test are presented in the last column of this table. For the item DEFECT the observed proportion of respondents who endorse this item is higher than the predicted proportion in the low score groups 2 and 3. (.49 versus .45, and .96 versus .94, respectively) whereas for the high score groups 4, 5, and 6 the reverse is true (.96 versus .98, .96 versus .99, and .98 versus .99). The HEALTH item, basically, behaves the same as DEFECT, i.e., the predicted proportion is higher than expected in

the low score groups 1 and 2, whereas in the high score groups the observed proportions are smaller than expected. The last column of Table C.5 presents the results of the one-sided binomial test. We see that for the items DEFECT and HEALTH, some of the differences between the observed and predicted proportions reach the one-sided significance level of 0.025. Since the differences between the observed and predicted proportions are systematic (positive differences in low score groups and negative differences in high score group) these results may indicate that the two items DEFECT and HEALTH do not discriminate enough compared to the other items in the analysis.

When we turn to the second and third page of table C.5, we see that for the RAPE, NOMORE, POOR, SINGLE, and ANY items the differences between the observed and predicted show a more or less random pattern. This is confirmed by the results of the binomial test for the five items which exceed all the one-sided significance level 0.025. So far, we may conclude that there is no reason to exclude the items RAPE, NOMORE, POOR, SINGLE, and ANY from a tentative (putative) Rasch scale.

Table C.6 shows the results of the binomial tests of the seven abortion items for the Germans. It is easy to see that for the DEFECT, HEALTH, RAPE, NOMORE, and POOR items several binomial tests are statistically significant ($p<.03$), whereas for the items SINGLE and ANY no difference between the observed- and predicted proportions reach this level of significance. For instance, the results of the binomial test for the item NOMORE show significant results

for the score groups 2, 4, and 5.. The pattern of the differences between the observed- and predicted proportions for this item shows that the observed proportion is smaller than the predicted proportion in the lower score groups, and for the higher score groups the reverse is true. This result indicate that NOMORE discriminates more than the other six abortion items. For the DEFECT, HEALTH, RAPE, and POOR items the deviations between the observed and predicted proportions are not systematic and, as a consequence, are more difficult to interpret. However, when we place these results in the context of our discussion "model fit and statistical testing" (see section 4.5), the significant bino-test results of these four items may be well be attributed to the relative large size of the German sample which makes the test very powerful for some score groups.

We complete the Rasch scale analysis for the Americans by looking at Molenaar's U_i test statistic which is presented for each item in Table C.5. This statistic tries to circumvent the problem of different power of the score groups, that is, the difference in the size of the score groups. Molenaar partitioned the sample a priori into three groups: two extreme high- and low score groups and into a middle class where no relation between the observed and predicted proportion is predicted (see Molenaar, 1983:60). Thus the U_i tests the extent to which the observed proportion of positive responses systematically deviates from the predicted proportion (Molenaar, 1983:63). The test statistic U_i has an approximate standard normal distribution under the null hypothesis that the item conforms to the

Rasch model. Large positive values of U_k indicate that an item trace line that is relatively flat; i.e, unrelated or negatively related to the remaining items. Large negative values of U_k are an indication that the item measures the same latent trait, but with a steeper trace line than the other items.

When we turn to Table C.5, we see that the results of the U_k statistic for the American confirms our earlier conclusions: For the items DEFECT and HEALTH, the U_k statistic has a value of 3.19 and 6.65, respectively, which is significant at the p=.05 level. In other words, the item trace lines of these items do not discriminate. For the items RAPE, NOMORE, POOR, SINGLE, and ANY, the values of the U_k statistic are not significant, and as a consequence, may be good candidates for a putative Rasch scale.

Table C.7 shows the values of the U_k statistic of the seven items for the German sample. The U_k-values range from -2.11 (RAPE) to 1.81 (HEALTH) which indicates that, (except for the RAPE item), the U_k values of the remaining items are not statistically significant. Thus we decide first to improve the general fit of the Rasch model to the German data by deleting the RAPE item from the analyses.

The results of the Fischer- and Schleibechner test of the seven abortion items for the Americans are presented in Table 8.3.

Table 8.3: Results from the Fischer and Schleibechner Test of the Seven Abortion Items for the American high/low Scoregroups

Pair-wise T-Tests of the equality of item parameters			
Variable	Item Parameters low	high	Difference low-high
Defect	-2.29	-1.10	-3.17
Health	-4.22	-1.80	-4.60
Rape	-2.57	-2.90	0.38
Nomore	2.40	1.08	5.38
Poor	1.47	1.03	2.03
Single	2.07	1.20	3.80
Any	3.15	2.48	2.33

Table 8.4: Results from the Fischer and Schleibechner Test of the Seven Abortion Items for the German high/low Scoregroups.

Pair-wise T-Tests of the equality of item Parameters			
Variable	Item Parameters low	high	Difference low-high
Defect	-3.21	-3.29	.11
Health	-3.54	-2.21	-2.83
Rape	-2.71	-1.70	2.76
Nomore	2.24	1.11	7.13
Poor	1.13	0.69	2.83
Single	3.10	2.60	2.83
Any	2.98	2.81	1.01

This ~~pairwise~~ analysis of the item parameter estimates shows that the items DEFECT, HEALTH, NOMORE and SINGLE vary considerably over the low and high score groups. Since we

decided earlier to interpret the results of the Fischer-Schleibechner test with some caution, we will first try to fit the Rasch model without the items DEFECT and HEALTH. When we look at Table 8.4, we see the results from the Fischer and Schleibechner test for the German sample. The results of the t-test show significant values for the items HEALTH, RAPE, NOMORE, POOR, and SINGLE. Again, because we do not only rely on the results of the Fischer and Schleibechner test, we decide to improve the general fit of the Rasch model to the German data by deleting the RAPE item from the scale.

In sum, the results of the first Rasch analysis of the seven abortion items for the Americans reveal clearly that the items DEFECT and HEALTH violate the assumptions of the Rasch model for the Americans. One important reason for the misfit of these two items is their low discrimination power in relation to the other items (see also Table 7.1.). For the German sample, however, it is not quite clear how we can improve the general fit of the model. The results of the Andersen test and Martin Löf test reveal significant results but the non-significant results of most of the the binomial tests and U_i values, do not provide much information on how to improve the fit of the Rasch model.

However, this result does not neccessarily imply that the hypothesis of the unidimensionality of the seven abortion items should be rejected. When we refer to the four axioms of the Rasch model, the sufficiency axiom is clearly violated (the discriminating powers of the items are not invariant across subgroups). However, since the sufficiency

axiom is violated the axiom of monotonicity is also violated, i.e., the tests based on a partitioning of the sum scores (Andersen Test, Martin Löf and BINO) do not reveal a strictly increase of the probability of a positive answer on *any* item as θ increases as specified by the logistic function of the item trace line. Thus an interesting alternative hypothesis for violation of these two axioms is the two-parameter logistic models that allows for a weighted sumscores as a sufficient statistic for the person parameter (see chapter 5).

Before turning to the two-parameter model, we complete our scale analysis for the Americans by looking at the fit of the Rasch model to the five remaining abortion items for the Americans. Table 8.5 reports the results of the Andersen and Martin Löf tests of the abortion items RAPE, NOMORE, POOR, SINGLE, and ANY for the American sample. Both test statistics show acceptable test results (T=27.63 with df=12 and X^2=7.63 with df=4, respectively) and thus do not give any reason to reject the Rasch model for the five items. Table C.7 shows the results of the binomial tests for each item and for each score group.

Table 8.5: Results from the Martin-Löf Test and Andersen Test of the Five Abortion Items for the Americans.

Martin-Löf-Test		
Score Group	Number of Observations	Contribution
1	343	8.47
2	98	5.68
3	98	4.95
4	88	8.54

T=27.63, df=12, p=.01
Results Andersen Test of equality of the five item parameters for low and high score groups:
X^2=7.63 df=4, p=.11

We see first, that none of the p-values of the five items are smaller than .025 (last column) which indicates that there are no systematic deviations between the observed and predicted proportions by the Rasch model. These results are confirmed by the non-significant values of the U, test-statistics which range from -1.83 for the item NOMORE to 1.58 for the POOR item. Finally, the results of the Fischer-Schleibechner test in Table C.8 indicate that the item parameter estimates of the items RAPE, POOR, SINGLE, and ANY do not differ significantly across the low- and high score groups. The NOMORE item seems to vary across the subgroups (T=2.74) according to statistical criteria. However, we will not elimate this item from the scale because first, the results of the Andersen- and Martin Löf

tests show a reasonably good fit; second, the results of the binomial tests and the U_i test statistic do not give any reason to delete the NOMORE item from the scale and third, since there are some problems with the Fischer-Schleibechner test the results of these test do not seem sufficient on this basis.

Van den Wollenberg (1979) argues that tests based on a raw score partitioning are sensitive to the axioms of monotonicity and sufficiency but, in certain circumstances are insensitive to the axiom of unidimensionality. Therefore, we also tested the unidimensionality axiom by using the Andersen Test. In this respect, the grouping was made according to the external criterion of sex. The results of these tests support very clearly our conclusion that the items RAPE, NOMORE, POOR, SINGLE, and ANY form a unidimensional Rasch scale for the American sample (Andersen Test, $X^2=5.4$ with 4 df).

For the German sample, we tried different possible combinations of scales and settled on a four-item scale which made up of DEFECT, RAPE, SINGLE, and ANY. Table 8.6 provides the results of the Andersen Test and the Martin Löf-Test for the four items.

Table 8.6: Results from the Martin-Löf Test and Andersen Test of the Four Abortion Items for the Germans

Martin-Löf Test		
Score Group	Number of Observations	Contribution
1	134	1.61
2	1017	1.27
3	268	10.74

T=13.62, df=6, p=.04
Results Andersen Test of equality of the five item parameters for low and high score groups:
X^2=6.5, df=3 ,p=.09)

Both statistics give statistically acceptable results (X^2=6.5 with 3 degrees of freedom and T=13.62 with 6 degrees of freedom) which indicates a considerably improvement of the degree of overall fit of the Rasch model compared to the test results of all seven items. The results of the binomial test are presented in Table C.9 for each item and for each score group. The p-values in the last column of the table indicate that for all four items, the difference between the observed- and predicted proportions are not significant (p >.025). In addition, the values of the test-statistic U_i range from -.051 for the item DEFECT to 1.949 for the RAPE item which represents an excellent fit for the German sample. The results of the Fischer and Schleibechner Test (see Table C.10) show that for the items RAPE, SINGLE, and ANY, the pair-wise t-test show values which are on the borderline of significance. For reasons mentioned above, we

will not consider these results as evidence for a possible misfit of the Rasch model to the four abortion items. Finally, as for the American sample, we also tested the unidimensionality of the DEFECT, RAPE, SINGLE, and ANY items by employing the Andersen Test. Again, the external criterion sex was used and the results strongly indicate the unidimensionality of the four items for the German sample (χ^2=1.54 with 3 degrees of freedom). Note that for the American as well as for the German sample, the final Rasch scales contain both at least one "medical" item. These results clearly do not support the conclusion drawn in earlier research that the abortion items measure two different dimensions: a scale made up of the four "difficult" items POOR, NOMORE, SINGLE, and ANY (abortion for social reasons) and the other scale made up of the three "easy" items HEALTH, DEFECT, and RAPE (abortion for medical reasons). [28]

Having established the unidimensionality of a subset of the seven abortion items, an important practical question is: how can we translate the results of the item- and person parameter estimates into the probability for a positive response according to the Rasch model? First note that if the Rasch model fits the data, we have established an interval scale. The mathematical units for the item and person parameter estimates defined by the Rasch model are called "logits". These units flow directly from the logistic

---

[28] We also tried to compute the Andersen Test for the overlapping items RAPE, SINGLE, and ANY of both scales using "country" as grouping variable. However, the program could not compute this test due to zero or perfect item scores in the groups.

model which specifies the estimated probabilities of a positive response as defined in equation (9) of section 5.3.3. When we rewrite (5.1), the probability of a correct response by person j to item k becomes:

$p_{jk}=\exp(\theta_j-\delta_k)/[1+\exp(\theta_j-\delta_k)]$

where θ_j is the estimated person parameter for respondent j and δ_k is the estimated difficulty parameter of item k. It follows that the odds for a positive response are

$p_{jk}/(1-p_{jk})=\exp(\theta_j-\delta_k)$

from which the natural log odds (to the base e=2.718) for a positive response becomes

$\ln[p_{jk}/(1-p_{jk})] = (\theta_j-\delta_k)$

These log odds are called "logits" and so the differences among items and persons are initially in logit units.

To see the usefulness of this result, suppose that two respondents have a value on the latent variable of θ_1, and θ_2, respectively. Then for item k, the log-odds are

$\ln O_{i1}=\theta_1-\delta_i$

and

$\ln O_{i2}=\theta_2-\delta_i$

for the two respondents, respectively. On substracting, we obtain,

$\ln O_{i1}-\ln O_{i2}=\theta_1-\theta_2,$

or $\ln (O_{i1}/O_{i2})=\theta_1-\theta_2.$

If the values of the latent variable differ by one point, i.e., $\theta_1-\theta_2=1$, then

$\ln (O_{i1}/O_{i2})=1,$

or alternately,

$O_{i1}/O_{i2}=\exp(1)$

= 2.718.

Thus a difference of one point on the scale of the latent variable, or equivalently, the log-odds scale, corresponds approximately to a factor of 2.72 in odds for a positive answer.

For instance, Table 8.7 shows the item parameters of the five items for the Americans. It is easy to see that the RAPE item is the easiest one (-3.88) and the item ANY is the most difficult (2.00). As mentioned before, the probability of a positive response in the Rasch model is a function of $(\theta-\delta)$.

Table 8.7: Item Parameter Estimates of the Five Abortion Items for the Americans.

Variable	Difficulty
Rape	-3.88
Nomore	0.80
Poor	0.34
Single	0.75
Any	2.00

Table C.12 gives examples of how these differences result in various probabilities on a positive answer for the items RAPE and POOR. Note that the more the person's parameter surpasses the item difficulty, the greater this positive difference and the closer the probability of a positive answer comes to one. For instance, the probability that a respondent with a person parameter value of 2.16 (sum

score of 4) will positively answer the item RAPE (item parameter value -3.88) is: exp(2.16+3.88)/[1+exp(2.16+3.88)] = 1.00 Further note that when a item is too "hard" for a respondent, the person parameter will be less than the item parameter value, their difference is negative and the person's probability of a positive answer is less than a half. For instance, when we turn to Table C.12, it can be seen that the POOR item is too difficult for respondents with a sum score of 1: the person parameter for respondents with a raw score of 1 is -2.29 and the item parameter of the POOR item is .34., their difference is (-2.29-.34)=-2.63 and the probability of a positive response is exp(-2.29-.34)/[1+exp(-2.29-.34)]=.07. Finally, table C.12 shows, that when the difference between the person parameter and the item parameter is small, the nearer the probability of a positive response comes to one half. Of course, the other Rasch results presented in this study can be interpreted in exactly the same way.

In sum, the results of the Rasch analyses show that RAPE, POOR, NOMORE, SINGLE, and ANY form a unidimensional Rasch scale for the American sample, whereas for the Germans, this scale contain the items DEFECT, RAPE, SINGLE, and ANY. Since both scales comprise both social items and at least one medical item, convergent evidence has been found for the unidimensionality of the abortion items (even though the items differ from those yielded by the Mokken test). Unfortunately, a direct comparison of American and West-German attitudes toward abortion is hampered by the fact that the Rasch scales do not contain the same items for

both countries and, as a consequence, are not comparable.

8.1 Results of the Logist Analyses

As mentioned earlier, the Mokken and Rasch models are the only scaling models that possess the feature of "specific objectivity". In particular in the present study, the properties derived from "specific objectivity" are important because valid comparisons of American and German abortion attitudes require that the measures of abortion attitudes are comparable across the two populations.

However in the two-parameter model, there is no sufficient statistic for the person parameter because a weighted sum rather than a simple sum of the item responses provides the information for the subject parameter. Therefore, the item responses cannot be reproduced from the scale scores since a respondent's θ value depends upon the particular pattern of item responses rather than the raw score. Thus, respondents with the same sum score can obtain different θ values if they endorse different items positively. This is highly impractical in the case of comparative research. For this reason, we discuss the two-parameter LOGIST analyses for illustrative purposes rather than for the main purpose of the present study: the establishment of a equivalent unidimensional scale of the abortion items for the United States and Germany.

In the following, we consider the results obtained from the joint maximum likelihood estimates of the item parameters under the two parameter logistic models for the Americans. By means of this analysis, we can test the

hypothesis that the discrimination power of the seven abortion items differ.

In Table C.13 we present the results of the two-parameter model for the seven abortion items for the American sample.[3'] The first column of this table shows the discrimination parameters of the seven items. The NOMORE item discriminates the most between respondents with liberal and more conservative attitudes toward abortion (2.4) whereas the discrimination power of the HEALTH item is extremely low (0.39). The third column of Table C.13 presents the difficulty parameter values δ for the seven items. It can be seen that the HEALTH item is the easiest item (-5.47) followed by the items RAPE (-1.41), DEFECT (-1.38), POOR (.76), SINGLE (.89), and ANY (1.76). Furthermore, Table C.13 also show the standard errors of the discrimination parameters (second column) and the standard errors of the difficulty parameters (fourth column). Immediately, the very high standard errors of both the discrimination and difficulty parameters of the HEALTH item warrant our attention (.4, and 31.2, respectively).

This result is in close agreement with the finding of Thissen and Wainer (1983) who state that the standard error of the difficulty parameter increases as the item gets more extreme. This problem arises when the sample size is not large enough to give acceptable estimates for these extremely located items. Furthermore, they point out that

---

[3'] The increase in the criterion function between two steps was less than .2% which means that the procedure converged and that the total change for all parameters approached zero.

standard errors of the discrimination parameters gets larger as slopes become more gradual.

Returning to Table C.13, we see that the discrimination parameter of the HEALTH item is the lowest compared to the other discrimination parameters and that, indeed, its standard error is relatively high (0.4). For the remaining items, the standard errors are still too high when we tolerate a maximum value of the standard error of .1 which is reasonable considering the size of the American sample (Thissen and Wainer, 1983).

Based on these results (and other not reported analyses) we conclude that, in order to get accurate parameter estimates of the abortion items under the two-parameter logistic model, both the American and West German samples are not large enough. (To get best results, we have to delete the "easy" items DEFECT, HEALTH, and RAPE from the scale. However, such a decision would not be based on important goodness-of-fit criteria but only on the drawbacks of the complexity of the estimation procedure of the two-parameter logistic model. Moreover, addition of items to the seven might increase the number of respondents at the extremes).

As we have mentioned earlier, the use of the LOGIST program is not without problems. One of the main difficulties of the program is that the joint maximum likelihood estimation procedure can yield parameter estimates with large biases and large standard errors of estimates even for a dataset of the size we use in the present study. A possible reason why this occurs is that we

do not know whether the item- and ability parameters are consistent and, as a consequence, we can only *assume* that it is possible to obtain standard errors of the maximum likelihood estimators of item and person parameters.

Another problem is, that the LOGIST program does not provide the user with an overall measure of the goodness-of-fit of the two-parameter model or with a measure of the goodness-of-fit for each item (for instance, graphical methods). Therefore, there is not much evidence accumulated in LOGIST to help the researcher judge whether or not the seven items form a unidimensional scale according to the assumptions of the two-parameter logistic model.

In sum, we may conclude that the two-parameter logistic model may be theoretically a more realistic model for assessing the unidimensionality of a set of attitudinal data than the more stringent one-parameter model. However, in order to get accurate estimates, the two-parameter model requires a data base that is larger than those typically found in survey research. That is, most surveys contain enough respondents to meet the criteria of the two parameter model but the number of items required is much larger than most survey researcher would tolerate. Moreover, the lack of any goodness-of-fit index, graphical method, or chi-square test makes it difficult to judge the unidimensionality of a set of items by the use of the LOGIST program.

8.2 Results of the NOHARM Analyses

As mentioned earlier, MacDonald (1980a, 1980b) and Hattie (1981) have suggested the use of non-linear factor analysis and the analysis of residuals as an alternative approach for assessing model-data fit. They argue that evidence of multidimensionality should be sought in a second- or higher-order indicator of misfit. In this respect, the magnitudes of the residual inter-item covariance matrix seem to offer a rational basis for a decision as to dimensionality.

Table C.14 presents the results of the item parameter estimates for the two-parameter model according to LOGIST and NOHARM for both samples. In order to get comparable results, the analyses are based on subsamples without those respondents who consistently reject or endorse all seven items. The first two colums of table C.14 presents the LOGIST estimates and the third and fourth colum of the same table show the NOHARM estimates. It can be seen that the parameter values obtained by the program LOGIST differ substantially from those obtained by NOHARM for both samples. For instance, the discrimination parameters *a* of the items in the first column estimated by LOGIST have all higher values than those estimated by the NOHARM method (third column).

Furthermore, the results for the Americans show for both programs that the discrimation power of the item NOMORE is the highest while the item HEALTH doesn't discriminate very well. Note, however, that the value of the discrimination parameter of the POOR item by LOGIST is lower

than the value of this parameter for the RAPE item (1.60 versus 1.72) whereas the estimates by NOHARM suggest that POOR discrimates better than RAPE (1.06 versus 1.02). The same can be said for the items ANY and DEFECT; whereas the discrimination parameter estimated by LOGIST show that DEFECT discriminates better than ANY (1.16 versus 1.06), it is less discriminating according to the estimates by the NOHARM method (.54 versus .52).

Table C.14 also shows that for the German sample, the item discrimination parameter of HEALTH is the highest by LOGIST followed by the items RAPE and ANY, whereas the item discrimination parameter of ANY is the highest by NOHARM followed by HEALTH and RAPE.

Furthermore, comparison of the difficulty parameters of both methods for both subsamples reveal, that for the items DEFECT, HEALTH, and RAPE, the estimates obtained by the LOGIST program (second column) are higher than those estimated by the use of NOHARM, whereas the converse is true for the items NOMORE, POOR, SINGLE, and ANY. Moreover, the program LOGIST provides a slightly different item ordering for both countries than the NOHARM method; HEALTH, RAPE, DEFECT, POOR, SINGLE, NOMORE, and ANY by LOGIST and HEALTH, DEFECT, RAPE, POOR, SINGLE, NOMORE, and ANY by NOHARM.

The differences in item parameter values produced by the two programs as well as the reversals in item difficulty ordering and discrimination power of the items, raise the question of whether the estimates are comparable. As mentioned before, the estimates by the LOGIST program are not very reliable. Therefore, we have to interpret its

results with some caution. However, we can also not accept the results by NOHARM without making some critical comments. First, inspection of the residual matrix for clusters of large residuals as well as the root mean square of these residual covariances of the items (which is an overall measure of misfit of the model to the data) revealed that the two-parameter model represents an excellent fit to the data. However, the same is true for the one-parameter model. For both models the root mean square residuals are considerably lower than the standard error of the residuals: for the one-parameter model the root mean square residual is 0.001 and the standard error is .15 whereas for the two-parameter model the root mean residual is 0.0005 and the standard error is also .15. Inspection of the residuals of both the one- and two-parameter model did not reveal any systematic deviations between the observed and fitted covariance matrices. However, since the discrimination parameters differ under the two-parameter model (see Table C.14), it is from a theoretical point of view impossible that the seven abortion items fit the Rasch model as well as the two-parameter model. Second, since there are reversals in the orderings of the difficulty levels estimates provided by LOGIST and NOHARM (of course, the item ordering based on p-values should be the same as the item ordering based on the difficulty parameters δ), NOHARM may not provide a suitable scaling method for the abortion data. Third, NOHARM does not directly estimate the θ values of the respondents in the samples to which they are fitted because it treats the latent trait (attitude) as a random variable.

8.3 Conclusion

An advantage of the NOHARM method is that it supplies a measure of the goodness of fit of the model in the familiar form of a residual matrix, and the sum of squares of its elements. However, an obvious limitation of the method is its assumption that the latent trait has a normal distribution. Violation of this strong assumption may be the main explanation for the ambiguous results we obtained by the use of this method. However, since this method is still in the exploratory stage of development, considerably more reseach is needed to assess whether the sum of absolute residuals is robust against non-normal distribution of the latent variable.[30]

In sum, the parametric latent-trait models do not provide a satisfactory solution for a meaningful comparison of attitudes toward abortion of Americans and West-Germans. The Rasch model in PML is preferable given the problems with LOGIST. However, the results of the Rasch analyses yield scales which are not robust across the two countries; i.e., the Rasch scales do not contain the same items for the Americans and the West-Germans. In addition, both PML and LOGIST do not provide estimates for respondents who endorse or reject all seven abortion items. For this reason, the American sample is reduced by 45% respondents, whereas for

---

[30]Table C.15. shows the difference between item parameters estimates of the one-parameter model with (N=1290) and without (N=702) the "zero" and "perfect" response patterns. The main effect of leaving those respondents in the analysis is that the scale shrinks due to the decreasing difficulty of the items. (Note, that there are considerably more respondents who approve all items compared to those who disapprove all items).

the German sample this number is 29%. It is needless to say, that the effect of deleting such a large number of respondents obscures the comparability of the results.

The NOHARM method has the advantage of providing item parameter estimates for the whole sample. However, the method lacks a good criterion for comparing nested model. That is, a comparison of the sum of the absolute residuals of the one and two-parameter model does not provide any information about which model best fits the abortion data.

9. SUMMARY AND DISCUSSION

9.1 Summary of Results

In the present study, we have applied some of the most commonly used latent structure models to the seven abortion items in the 1982 NORC GSS and West German Allbus combined files. The research questions were: First, do the seven abortion items constitute a unidimensional set of items for each country, separately. Second, do the set of selected items or a smaller subset comprise a scale that is robust across the populations under study. Third, if we find a robust scale or scale(s), which country is more accepting of legalized abortion, the United States or West Germany? Finally, how we can use the metric scale(s) developed in the course of scale analysis as a "reference" indicator in cross-population LISREL models.

The results of our analysis showed; a) convergent evidence for the unidimensionality of the abortion items; b) Americans are slightly more liberal than West-Germans in their attitudes toward abortion; and c) variation in attitudes is greater for Americans than for West-Germans.

Convergent evidence of unidimensionality of the abortion items

We used several methods to assess the unidimensionality of the seven abortion items in both countries, respectively: The Mokken scale analysis, latent class analysis, the Rasch model, and Birnbaum's two-parameter model. All these models provided some convergent evidence of the unidimensionality

of the abortion items. However, the number of items that comprises the unidimensional scales appeared to be dependent on the pecularities of the specific model.

Using the non-parametric Mokken methods, we found that the seven abortion items constitute a unidimensional scale in both countries and that the items DEFECT, POOR, NOMORE, and SINGLE form a scale that is robust across the two populations.

By employing the latent class models, further convergent evidence of the unidimensionality was found for the four robust Mokken items. Goodman-type models proved to fit the data very well for both countries.

Next, we considered the parametric Rasch model to detect more specific measurement properties of the abortion items for both countries. The results of these analyses by the computer program PML revealed that the items RAPE, POOR, NOMORE, SINGLE, and ANY formed a unidimensional Rasch scale for the American sample, whereas for the Germans, this scale contained the items DEFECT, RAPE, SINGLE, and ANY.

Both the Rasch and Mokken model satisfy each in its own way, the criterion of specific objectivity (the most important property of social measurement). However, the selected items that form the unidimensional scale are different for both methods.

Notwithstanding these differences, the results still support the unidimensionality of the abortion items. Since, regardless of the method used, the different scales contain both medical and social items. A direct comparison between the resulting Mokken scale and Rasch scales is hampered

because the Mokken and Rasch scales do not yield the same set of items.

As an alternative strategy, we applied LOGIST and NOHARM to the abortion items to assess unidimensionality. We investigated whether the two-parameter model would be more realistic than its one-parameter counterpart. The LOGIST analysis revealed that the two-parameter model may not be a good alternative to the one-parameter model since the standard errors were very large and no criterion of goodness-of-fit was provided. Application of the NOHARM method showed that both the one and two-parameter model fit the data very well. Due to lack of proper goodness-of-fit criteria, the results of the LOGIST and NOHARM analyses seem to be inappropiate for the establishment of a well founded unidimensional scale for the seven abortion items.

Americans are more liberal in their attitudes towards abortion than West Germans

The metric of the newly established unidimensional Mokken scale has been assessed with the log-multiplicative method ANOASC. The item "church attendence" proved to be a suitable instrumental variable with which we could successfully recalibrate the ordinal abortion attitude scale scores for both countries. A comparison of Americans and West-Germans on this metric scale showed that Americans are somewhat more accepting of legalized abortion. Validation of this result by the use of different instrumental variables provided convergent evidence of the difference between the means of the abortion scale scores for the Americans and

West-Germans; i.e., Americans are slightly more likely to approve of legal abortion.

Variation in attitudes is greater for Americans than for West-Germans

Results of the F-test for differences between the variances of the calibrated scale scores show that the variance is greater for Americans than for West-Germans ($p<.01$). First, Americans hold proportionately a more liberal view with respect to the social items, whereas proportionately more West-Germans endorse the medical items. Moreover, the variance is somewhat greater for the Americans than for the West Germans. Second, the distribution of the scale scores shows the same picture; i.e., the percentage of cases in *both* extreme categories is greater for the Americans. Third, the distances between the extremes of the metric scales by CHATTEND are greater compared to the distances between the intermediate scale scores for both countries.

We find that the calibrations of the categories are sensitive to the particular instrumental variable chosen. However, all calibrations provided convergent evidence of the difference in variation in attitudes between the Americans and West-Germans; i.e., the variation in attitude toward abortion is greater for the Americans than for the West-Germans.

We conclude by arguing that only a comparison of *scale* scores suggest that the variances and the means of the abortion scale scores differ for the Americans and

West-Germans. Looking at individual items do not provide that information since comparison of item marginals confounds differences between Americans' and West-Germans' positions on the latent variable with the cross-national differences between the *Items'* positions on the latent variable. For this reason, we argue that the two-dimensional solution of the abortion items is an artifact of the difference in variation of attitudes toward abortion for the Americans and West-Germans.

9.2 Discussion

Review of the literature revealed that opinions on the seven abortion items differ in both the substantive literature on abortion, and the methodological literature that uses these items to discuss unidimensionality. At issue is whether the seven items constitute one or two dimensions: one which contains the medical or "hard" reasons (HEALTH, DEFECT, and RAPE), the other which contains the "social" or "soft" reasons (POOR, SINGLE, NOMORE, and ANY).

The major empirical difference between the two sets of items is that the percentage of respondents who approve of abortion for medical reasons is substantially greater than the percentage who approve of abortion for social reasons. Put another way, the medical items are less "difficult" than the social items. However, whether this difference justifies the separation of the seven items into two different measures is another question.

With regard to the comparison of abortion attitudes between the two countries, we noted the problem of

seperating the difference in general attitudes, as measured by the *set* of items, from the difference in the difficulty of *specific* items. In terms of Berry's (1980) discussion of measurement in cross-cultural settings, it is possible that the abortion items lack *psychometric* equivalence across the American and West German population. For this reason, it becomes important to decide whether the scalability of a set of items is the same for different populations. In other words, we need to develop unidimensional measures of the limits of abortion approval and to keep difference in scale unidimensionality and difference in item difficulty seperate. Only then we can make valid comparisons between the abortion attitudes of the Americans and West Germans.

In the present study, we approached these problems of unidimensionality and comparability by the application of both latent trait and latent class models to the seven abortion items. Based on the results of the different methods, we tried to find convergent evidence that the abortion items constitute a single dimension. In addition, by using this variety of models, we were able to examine the applicability of these methods to a typical attitudinal data set.

For discussing the models, we will focus on the following issues: 1. assumptions of the models; 2. assessment of goodness of fit criteria; and 3., sample size needed to get accurate parameter estimates.

Assumptions of the Models--

The non-parametric Mokken model makes no assumption about the functional form of the item trace lines. For this reason, we refer to this model as non-parametric, and the resulting scale scores and item difficulties constitute ordinal scale scores. The only constraints that the Mokken model puts on the item trace line was referred to as the assumption of double monotony, which implied that the ordering of the items is specifically objective. The latent class model can also be considered as a non-parametric model because it does not impose a functional form on the item trace lines, nor does the model require the assumption of normality of the latent variable. The latent class probabilities and the conditional probabilities are the two fundamental parameters of the model.

Instead, the parametric latent trait models make rather strong assumptions about the functional form of the item-trait relation. With respect to the one-parameter Rasch model, the model is specified by one-parameter logistic function. It is assumed that all items have equal discrimination power. The assumptions of the Rasch model are highly restrictive and appear to be rarely met in social science research. For this reason, the two-parameter model is considered by many as a preferable alternative to the one-parameter model. The two-parameter model is less restrictive than the one-parameter model in that it allows for varying discrimination power of the items. As a consequence, the model does not have the property of specific objectivity.

Finally, the NOHARM method is used to estimate the one and two-parameter models. In contrast to the other latent trait and latent class models, the NOHARM method assumes that the latent trait has a normal distribution.

Assessment of goodness-of-fit criteria

The Mokken model provides statistical tests which allow the researcher to judge the fit of the model. By means of these tests we can, first, test the null hypothesis that the population values of the coefficients of scalability are equal to zero or some larger cutoff value. Second, one can use a statistically based procedure for adding or deleting items from a putative Mokken scale. Finally, one can test the null hypothesis that the pattern of relationship among the items holds across two or more populations. With respect to the latent class models, maximum likelihood estimators of the conditional and latent class probabilities are provided by the computer program MLLSA. The fit of the latent class models can be tested with the chi-square test statistic.

Turning to the parametric models, the Rasch model in PML allows the derivation of fit statistics with known distributional properties such as, the Andersen test and the Martin-Löf test. The CML procedure employed in PML, uses the total score as a sufficient statistic. Therefore, item parameters can be estimated independently of person parameters. Instead, the UML procedure is used for the estimation of the parameters of the two-parameter model by the computer program LOGIST. This procedure may not provide consistent parameters estimates due to simultaneously

estimation of the item and person parameters. Therefore, fitting the two-parameter model has presented problems of estimation for which no clear and universally accepted solutions have been found. LOGIST, does not provide the user with any measure of the goodness-of-fit of the items to the data.

The NOHARM method provides the root mean square of the residual covariances of the items as an overall measure of the fit of the model to the data. However, this index cannot be used to compare nested models; i.e., the selection of the best-fitting model by NOHARM is almost impossible.

Sample size needed in order to get accurate parameter estimates

The Mokken method is successfully used in research on the unidimensionality of psychological variables in experiments that contain 6 to 13 measures of child development and 22 to 178 subjects (Kingma, 1984; Kingma & Loth, 1985; Kingma & Reuvekamp, 1984; Kingma & Ten Vergert, 1985). Thus, the Mokken methods is an adequate latent trait model for most attitudinal data where the number of items is often not more than, say, 10. With regard to the latent class models, moderately sized item pools (4 items) and very large sample sizes are needed to avoid too many empty cells.

An advantage of the conditional maximum likelihood in ~~PML~~ is that the method does not need a large number of items. However, the model tests used in PML require large sample sizes: If the model fits, parameters obtained in one subgroup will be the same as those obtained in any other

subgroup. However, these tests are rather unstable when the number of respondents in a subgroup is low. Since most attitude surveys contain enough respondents to meet this requirement, and since the number of items is often relatively small, the computer program PML is well able to cope with attitudinal data. In contrast, the two-parameter model estimated by LOGIST requires data bases that are much larger than usually found in survey research. When evaluating "the fit" of the abortion items, we found that the standard errors of the parameter estimates were too large to provide accurate estimates of the items parameters under the model. The main reason why this occurred was that the effective size of both the American and West-German sample was too small to provide acceptable estimates for the extremely located items. Of course, this problem is also related to the population distributions of the abortion items in both countries.

Finally, due to the newness of NOHARM, research on the robustness of the method is lacking which makes the evaluation of this method difficult.

Our evaluation of the application of the Mokken methods, the Rasch model and Birnbaum's two-parameter model to the seven abortion items was focussed on four issues that caused serious problems in the course of this study. The Mokken methods introduced in this study was able to deal well with all these issues. First, it is not driven by the often unrealistic strong assumptions of the logistic parametric models or the assumption of normality of the latent variable as in the NOHARM methods. Second, the Mokken

scale analysis provides a sound statistical basis to determine the unidimensionality and robustness of a set of items. With respect to the parametric models, only the tests of fit derived from the conditional maximum likelihood procedure for the one-parameter model provide statistically sound test of fit. However, it became clear that none of these statistical tests of fit were alone sufficiently powerful to warrant the conclusion that the data fit the model. For this reason, other criteria in evaluating model fit were used in addition to the statistical tests. Because the computer program PML provides both statistical tests as well as other goodness-of-fit indices (for instance, graphs, BINO-tests and residuals) this program is superior to programs like LOGIST and BICAL with respect to the assessment of fit for the one-parameter model. Fourth, the Mokken model and the NOHARM method both retain the respondents who approved or disapproved all seven abortion items. As this study showed, deletion of those respondents from the scale analysis by the programs LOGIST and PML, obscured the nature of the sample and, as a consequence, generalizations to the populations cannot be made. We argued that, although most research reports on attitude measurement do not refer to this problem, the high proportion of respondents who rejected or accepted all seven items was not atypical for survey research in general.

Fifth, in contrast to the logistic models, the Mokken method (and probably the NOHARM method) does not require a data base that is larger than those typically found in survey research. In particular, the number of items required

is much larger than most survey researchers would tolerate.

Latent class analyses

In the present research, a distinction was drawn between two different traditions in latent structure analysis which we termed latent trait and latent class analysis. It was noted that both traditions were unified by the fact that both take local independence as the basis for measuring the latent variable. So far, our discussion has been focussed on the applications and results of the latent trait analyses. In the following we will discuss the use of the latent class scaling models to the four robust abortion items for both countries.

Although one can find substantive applications of latent class models in stratification research, their use in scale analysis is largely limited to methods papers. Therefore, we considered it important to examine the applicability of the latent class scaling models to a typical attitudinal data set which has been used in different substantive inquiries. Another purpose of the present study was, to compare the results of the latent trait and latent class analyses of the *same* abortion data in order to find convergent evidence that the seven items constitute a single scale.

This question is important, since there is a fundamental difference between the two methods which stems from the way in which the models express the relationship between the response probabilities and the latent variable. The fit of the four items to the latent trait Mokken model

was interpreted as evidence that these items could together be considered as dominated by a single dimension. In contrast, the latent class models do not test the relationship between the response probabilities and the latent variable. That is, the model assigns proportions of respondents to the latent classes which give best fit to the data but does not test the assumption that the latent trait is distributed on just these few points or that the latent classes are ordered on a line.

In addition, latent trait models differ from latent class models in that the former posit a continuous latent variable, while the latter posit a discrete latent variable. Furthermore, an important technical difference between the two traditions is that most latent trait analyses require only the bivariate cross-classifications of the items or the item variances and covariances, while latent class analyses typically require the full n-way table, i.e., higher order moments (Mooijaart, 1982). Our study revealed that, in order to avoid too many empty cells, a very moderately sized item pool and a very large sample size were needed. Even with the relatively large sample sizes and the moderate number of items we used in the present study, there were too many empty cells to compute the fit of the models.[3'] Due to this problem, we decided to analyze the latent class scaling models for the four robust items (DEFECT, POOR, NOMORE, SINGLE) and to compare the results of these analyses with the results of the Mokken test. Thus knowing that the items

---

[3'] For the same reasons, log-linear Rasch is highly impractical for the analysis of more than four items (Duncan, 1984b; Kelderman, 1984).

formed a unidimensional and equivalent scale, we used the responses to these four items to assign respondents to the latent classes of the scaling models.

As mentioned before, the results of the latent class analyses revealed that Proctor's model and unscalable class represented the data very well for the Americans, whereas the "pure" Goodman model was an appropiate scaling model for the West German sample. By means of the simultaneous latent structure models we were able to test the difference between the distributions of the latent variable. The results of this test revealed, that the distributions of the latent variable "attitudes toward abortion" was not the same across the United States and West-Germany.

Another issue we want to discuss is, that for both countries, the percentage of members of the unscalable class could hardly be reduced by the introduction of more flexible error-rate models and unscalable class. Since we already suspect that the four items do form a robust unidimensional scale, this finding may be an indication that the possibilities of modelling the error structures in latent class scaling models is not flexible enough. In addition, the program does not allow us to specify inequality constraints on the error rates. These inequality contraints follow from the assumption of double monotony for the Mokken and Rasch models. Nevertheless, we suggest that it should apply to the conditional probabilities if the latent classes form a single scale. Finally, the models are often underidentified when the probabilities of a false-positive and a false-negative are allowed to differ for each item

("latent distance type" restrictions). These limitations of the method make model-data fit extra difficult and, in addition, make the comparison between the latent class procedures and latent trait techniques more complex.

Assigning a metric to ordinal scale scores

We refined the ordinal robust scale, by assigning a metric to the categories of the Mokken scale according to the log-multiplicative association method discussed by Clogg (1982, 1984a, 1984b) and Goodman (1984, 1987). Clogg (1982, 1984b) assigned a metric to both the Guttman scale types as well as to the category of respondents whose responses were not in accordance with the scale types. Because the Guttman model makes no allowance for measurement error, Clogg was forced to develop an ad hoc procedure by condensing the error response patterns into one category. However, it may be argued that assigning scale scores to a class or category of respondents who do not respond conform the Guttman scale types is dubious because, according to the Guttman model, the responses of the members of this category are assumed to be not scalable. As a consequence, they do not do not contain useful information for the construction of an unidimensional scale or the establishment of a metric scale.

In the present study, we have tried to circumvent these problems by extending Clogg's method of assigning a metric to deterministic Guttman type scales to unidimensional scales selected using Mokken's non-parametric latent trait model. The Mokken model (as well as the other latent trait models) try to select those items on which the *whole*

population is scalable. In contrast, the deterministic models use the entire set of items in the analysis and assigns error to *cases* that do not satisfy the Guttman scale types. Because the Mokken model is a stochastic model, the possibility of measurement error is built into the model by positing a probabilistic relation between the values of the latent variable and the probability of a positive response to the item. As a consequence, our approach avoids the problem of assigning respondents to an "error category". Moreover, because we substituted metric values for the ordinal scale values of the robust scale, we felt confident that no items were included that might not be robust across the populations being compared.

By using more than one instrument variable, we were able to test the robustness of the rescaling for the American sample. The results of these tests indicate that the estimates of the distances depend on the instrument variables chosen. However, we still accept Clogg's conclusion "that these methods are a clear improvement over the popular 'method' of assuming that ordinal variables are interval level with equal intervals" (Clogg, 1984:222). As this study shows, the assumption of equal intervals between the scale values of the equivalent scale would be entirely inappropiate.

9.3 Suggestions for Future Research

The virtues of latent trait theory and the potential it holds for solving hitherto unsolvable problems in the area of attitude measurement makes this approach invaluable to

social researchers. The basic conception of latent trait theory is that the measurement features of an item are defined by its item characteristic function and that the measurement scale must not depend on the individual or group being measured. However, it would be improper to leave the reader with the impression that latent trait theory is a measurement panacea. As this study showed, to enjoy the advantages of the theory, we must be ready to pay.

Most latent trait models are based on strong assumptions that forces us to follow certain statistical constraints which in the end may invalidate the results. In this respect, the largest hurdle in the application of latent trait models to attitudinal data is to test the validity of those assumptions.

The Mokken method as screening method

A suggestion for future research is to use the non-parametric Mokken method as a preliminary step in the construction of unidimensional scales and the accompanying Mokken test as a means for developing equivalent attitude scales. The Mokken method of scale analysis provides a test of unidimensionality that incorporates the stochastic nature of social measurement as part of the measurement model. It also provides a test for assessing the equivalence of the scale scores across different time periods and cultural setting. Finally, it also can be used to see whether new items are unidimensional with respect to some identified scale.

However, the main advantage of the Mokken method over the other latent trait methods is the balance it strikes between flexibility and rigour. On the hand, it provides a distinct improvement over the *ad hoc* construction of Likert scales or inappropiate use of factor analysis that abound in previous studies of abortion attitudes. On the other hand, it is more practical to use than the other parametric latent trait models.

Design of better questionnaires

Another suggestion for future research is the need to design better questions. In particular, in measuring abortion attitudes we need items that are more difficult. The introduction into 1982 NORC GSS's of the ANY item is a step in this direction. Another possibility is, to ask whether the respondent would approve of abortion because the fetus was the wrong sex. Of course, the problem that addition of new items to an existing scale poses is whether these new items destroy the unidimensionality of the scale. In this context, we note again the value of the Mokken method since it provides procedures for addressing this question.

While the need to design better questionnaires is important, it nonetheless does not garantee that the item difficulties are the same across populations. From a comparative point of view, we emphasis the need to keep seperate differences in item difficulty of specific items and differences in the average value of the latent trait. The first presents change in the respondent's average

position on the latent trait measured: with respect to abortion, the willingness of people to support legal abortion on any conceivable grounds, whether covered by the items in the scale. The second represents change in the position of the items on the latent trait: the willingness of all people to support legal abortion on the specific grounds stated by the items, regardless of the general attitude toward abortion. These two sources of difference can operate relatively independently of one another and we need scaling methods to seperate the two.

The need for standardization of computer programs

Latent trait theoretic procedures are mathematically complex and its applicability is almost totally dependent on the availability of large computers. Therefore, the choice of models is often based on practical considerations, such as the availability of computer programs and local expertise.

An interrelated issue here is that the item parameter procedures are inextricably intertwined with the computer software that implements the estimation procedure. For instance, the LOGIST and BICAL programs have implemented the JML procedure, the PML program the CML procedure, NOHARM the least-squares methods, and the program BILOG is suitable for marginal maximum likelihood estimation and Bayesian modal estimation of the item parameters. Moreover, as the program gets more complex, the number of analysis options increases which also influence the manner in which the analysis are performed and hence the results. The same is true for the

differences between the metrics of the parameter estimates which make a routinely comparison of parameter estimates across studies extra difficult. A suggestion for future research is that each paper should specify the options used in each computer program employed, the metric of the parameter estimates, and the parameter estimation procedure implemented in the program.

Another problem is, that many articles dealing with latent trait models employ computer programs that have not been released for general use, or that are still in the initial stages of development. As a consequence, the documentation for general distribution about these programs is often not accessible. Due to the shift toward micro computers, this tendency of diversification of different programs and algorithms will be greatly enlarged. Therefore, in future research, we suggest that the listings of the programs used are attached as an appendix to the articles. Such a publishing approach may results in a wider application of these computer programs and will improve the discussion about the algorithms used.

Assessment of goodness-of-fit.

At present, there are only statisticially justifiable procedures for the assessment of model fit are available for the Rasch model when *conditional* estimators if item parameters are obtained. As noted before by Lord, local independence follows automatically from unidimensionality. However, what to take as evidence for multidimensionality, in particular in the case of the two- and three parameter

models is not clear yet.

A suggestion for future research could be, the development of a second-order statistic (like Van den Wollenberg's Q_2 for the one parameter model) for the two- and three parameter model. This statistic could circumvent the problems of the first-order indicators of misfit and seems to offer a rational basis for a decision as to dimensionality. The potentially promising method NOHARM is one the most appropiate programs to implement this second-order indicator of misfit.

LISREL model with dichotomous and continuous variables

To illustrate the value of the metric scale for use in cross-national LISREL models, we attempted to analyze the model as shown in Figure C8.1.

The basic idea was to use the metric scale as the reference variable ABORT. By doing this, the metric of this scale would establish the metric of the latent variable. In addition, the estimates of the factor loadings of the other items HEALTH, RAPE, and ANY would also be expressed in terms of the reference variable ABORT. In order to find convergent evidence that the means and variances differ between the United States and West Germany, we intended to examine similarities and differences in means and variances of the two countries using the LISREL model.

We started by computing the polyserial and tetrachorical correlations: first, because the application of factor analysis to the dichotomous items may lead to difficulty factors (see Section 1.2), and second, because

the model consists of a mixture of dichotomous and continuous items. Assuming that the observed dichotomous and continuous variables are manifestations of underlying latent variables which are normally distributed, the polyserial correlation coefficient is the (latent) correlation between a pair of latent response to the variable (Muthén, 1983, 1984).

Recent versions of the LISREL computer program include algorithms for computing tetrachoric and polyserial correlations. However, the program does not provide procedures to handle this feature in combination with stacked models. The problem is that the latent variable is standardized and, so, the metric is lost. As a consequence, the question reoccured whether the slope of the reference variable ABORT is invariant across the two countries.

One suggestion to circumvent this problem is to use only continuous variables for the use of cross-national LISREL models and to analyze their covariance matrix. An example how this can be accomplished is given in Figure C8.2. Again the item ABORT is used as the reference variable. However, the items SUICIDE and EUTHENASIA are assumed to be continuous. By means of this model we can examine whether the attitudes toward abortion are part of a more general concept, e.g., "prolife" attitude.

By analyzing the covariance matrix the scale of the latent variable and the observed variable is not standardized. This means that the reference variable ABORT will prove useful in establishing the metric of both the latent variable and the other variables for the Americans

and West Germans. As a consequence, meaningful comparison can be made between the means of the latent variable for the two countries. Unfortunately, we were not able to test the latter model to the data because the German data set does not provide the variables that were relevant for the model.

Another solution for the problem mentioned above, might be the use of the computer program LISCOMP (Muthén, 1987) which is especially well suited for categorical and non-normal data.

Non-linear factor analysis

In future research non-linear factor analytic methods should be examined. These methods have been developed for dichotomous items and assume that the latent trait is normally distributed. Since these methods provide promising indices that can be used as decision criteria for determining unidimensionality, research on the robustness of these methods (particularly when there is a non-normal distribution of the latent trait) is needed.

Latent Class Methods

Of latent class analysis and latent trait models, the latter strikes us as most promising in the context of attitude measurement. However, latent class scaling models have not yet been applied on a broad scale in the area of attitude measurement and, therefore, constitute a less well developed alternative to the latent trait models. Moreoever, there is not much information about the relationship between latent-class and latent-trait models (exceptions are Clogg,

1987; Van der Linden, 1978). Research needs to be done to deal with restrictions on response probabilities so that the results of both approaches are easier to compare. Of course, this research involves the problems of model identification and parameter estimation for underidentiefied models. Another suggestion for future research is, the development of appropiate data analysis procedures for handling moderate item sets. Researchers that attack this problem will facilitate the use of latent class models in the area of survey research.

Bibliography

Allen, M.J., & Yen, W.M. (1979). Introduction to measurement theory. Monterey: Brooks/Cole Publishing Company.

Andersen, E.B. (1972). "The numerical solution of a set of conditional estimation equations". The Journal of the Royal Statistical Society: 42-54.

Andersen, E.B. (1973). "A goodness of fit test for the Rasch model". Psychometrika 38: 123-140.

Andersen, E.B. (1977). "Sufficient statistics and latent trait models". Psychometrika 42: 69-81.

Anderson, A.B., A. Basilevski and Hum, D.P. (1983). "Measurement: Theory and techniques", pp.231-287 in P.H. Rossi, J.D. Wright and A.B. Anderson eds., Handbook of survey research . New York: Academic Press, Inc.

Andrich, D. (1978). "Relationship between Thurstone and Rasch approaches to item scaling". Applied Psychological Measurement 3: 449-460.

Andrich, D. (1982). "Using latent trait measurement models to analyse attitudinal data: A synthesis of viewpoints", pp.89-126 in D. Spearitt, ed., The improvement of measurement in education and psychology. Contributions of latent trait theories. Hawthorn: Australian Council for Educational Research.

Andrich, D. (1985). "An elaboration of Guttman scaling with Rasch models for measurement measurement", pp.33-80 in N.B. Tuma, ed., Sociological methodology 1985. London: Jossey-Bass Publishers.

Arney, W.R., & Trescher, W.H. (1976). "Trends in attitude toward abortion, 1972-1975". Family Planning Perspectives 8: 117-124.

Barton, M.A., & Lord, F.M. (1981). "An upper asymptote for the three-parameter logistic item-response model". Research Bulletin 81-20. Princeton, NJ: Educational Testing Service.

Beniger, J.R.. (1984). "Mass media, contraceptive behavior, and attitudes on abortion: Toward a comprehensive model of subjective social change", pp. 475-500 in C.F. Turner and E. Martin (Eds.). Surveying Subjective Phenomena, Vol. 2. New York: Russell Sage Foundation.

Berry, J.W. (1980). "Introduction to methodology", in J.W. Berry, ed., Handbook of cross-cultural psychology, (Vol.2). London: Allyn and Bacon, Inc.

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability", pp.397-497 in F.M. Lord and M.R. Novick, eds., Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Blalock, H.M. (1982) Conceptualization and measurement in the social sciences, Beverly Hills: Sage Publications Inc.

Bohrnstedt, G.W. (1983). "Measurement", pp.70-121 in P.H. Rossi, J.D. Wright and A.B. Anderson, eds., Handbook of survey research. New York: Academic Press.

Campbell, D.T. (1950). "The indirect assessment of social attitudes". Psychological Bulletin 47: 15-38.

Campbell, N.R. (1921). What is science? New York: Dover Publications, Inc.

Campbell, N.R. (1928). An account of the principles of measurements and calculations . London: Longman, Green.

Clogg, C.C. (1977). "Unrestricted and restricted maximum likelihood latent structure analysis: a manual for users". Working Paper 1977-09, Pennsylvania: Population Issues Research Office.

Clogg, C.C. (1979). "Some latent structure models for the analysis of Likert-type data". Social Science Research 8: 287- 301.

Clogg, C.C. (1981). "New developments in latent structure analysis", pp.215-246 in D.M. Jackson and E.F. Borgatta, eds., Factor analysis and measurement in sociological research. Beverly Hills: Sage Publications.

Clogg, C.C. (1982). "Using association models in sociological research: Some examples". American Journal of Sociology 88: 114-134.

Clogg, C.C. (1984a). "Some statistical models for analyzing why surveys disagree", pp.319-366 in C.F. Turner and E. Martin, eds., Surveying subjective phenomena. New York: Russell Sage Foundation.

Clogg, C.C. (1984b). "Using association models in sociological research: Some examples", pp.203-223 in L.A. Goodman, ed., The analysis of cross-classified data having ordered categories. Cambridge, Mass.: Harvard University Press.

Clogg, C.C. (1987). "Latent class models for measuring", in press in R. Langeheine and J. Rost, eds, Latent trait and latent class models. New York: Plenum.

Clogg, C.C., & Goodman, L.A. (1984). "Latent structure analysis of a set of multidimensional contingency tables". Journal of the American Statistical Association 79: 762-771.

Clogg, C.C., & Goodman, L.A. (1985). "Simultaneous latent structure analysis in several groups", pp.81-110 in N.B. Tuma, ed., Sociological Methodology 1985 . San Francisco: Jossey-Bass.

Clogg, C.C., & Goodman, L.A. (1986). "On scaling models applied to data from several groups". Psychometrika 51: 123-135.

Clogg, C.C., & Sawyer, D.O. (1981). "A comparison of alternative models for analyzing the scalability of response patterns", pp.240-280 in S. Leinhart, ed., Sociological Methodology 1981. London: Jossey-Bass Publishers.

Combs, M., & Welch, S. (1982). "Blacks, whites, and attitudes toward abortion". Public Opinion Quarterly 46: 510-520.

Coombs, C.H. (1964). A theory of data. New York: Wiley.

Cronbach, L.J. (1951). "Coefficient alpha and the internal structure of tests". Psychometrika 16: 297-334.

Davis, J.A., & Smith, T.W. (1985). General Social Surveys, 1972-1985: Cumulative Codebook. Chicago: National Opinion Research Centre.

Dawes, R.M. (1972). Fundamentals of attitude measurement. New York: Wiley.

Dayton, C.M., & Macready, G.B. (1976). "A probabilistic model for validation of behavioral hierarchies", Psychiometrika 14: 189-204.

Dayton, C.M., & Macready, G.B. (1980). "A scaling model with response errors and intrinsically unscalable respondents". Psychometrika 48: 343-356.

DeGruyter, D.N.M., & VanDerKamp L.J.Th. (1984) Statistical models in psychological and educational testing . Lisse: Swets & Zeitlinger.

Duncan, O.D., Sloane, D.M., & Brody, C. (1982). "Latent

classes inferred from response-consistency effects," pp.19-64 in K.G. Jóreskog and H. Wold, eds., Contributions to economic analysis. Amsterdam: North-Holland Publishing Co.

Duncan, O.D. (1984a). Notes on social measurement: Historical & critical. New York: Russell Sage Foundation.

Duncan, O.D. (1984b). "Rasch measurement: Further examples and discussion", pp.367-403 in C.F. Turner and E. Martin, eds., Surveying subjective phenomena, (Vol. 2). New York: Russell Sage Foundation.

Ebaugh, H.R., & Haney, C.A. (1980). "Shifts in abortion attitudes: 1972-1980". Journal of Marriage and the Family 42: 491-499.

Edwards, A.L. (1953). Techniques of attitude scale construction. New York: Appleton-Century-Crafts, Inc.

Fienberg, S.E. (1983). The analysis of cross-classified categorical data. Cambridge: The MIT Press.

Finlay, B.A. (1985). "Correlates of abortion attitudes and implications for change", pp.178-190 in P. Sachdev ed., Perspectives on abortion. London: The Scarerow Press, Inc.

Fischer, G.H. (1974). Einfuehrung in die Theorie psychologischer Tests. (Introduction to psychological test theory). Bern: Huber.

Fischer, G.H., & Schleibechner, H.H. (1970). "Algorithmen und Programmen für das probabilistische Testmodel von Rasch". Psychologische Beiträge 12: 23-51.

Fraser, C. (1980) NOHARM. An IBM PC Computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: Centre for Behavioral Studies, The University of New England.

Glass, J., Bengston, V.L., & Dunham, C.H. (1986). "Attitude similarity in three-generation families: socialization, status inheritance, or reciprocal influence"? American Sociological Review 51: 685-698.

Green, B.F. (1954). "Attitude Measurement", pp. 335-369 in G. Lindzey, ed., Handbook of social psychology. New York: Addison-Wesley.

Goodman, L.A. (1974a). "Exploratory latent structure analysis using both identifiable and unidentifiable

models". Biometrika 61: 215-231.

Goodman, L.A. (1974b). "The analysis of qualitative variables when some of the variables are unobservable Part I-A modified latent structure approach". American Journal of Sociology 79: 1179-1259.

Goodman, L.A. (1975). "A new model for scaling response patterns: An application of the quasi-independence concept. Journal of the American Statistical Association 70: 755-768.

Goodman, L.A. (1979a). "The analysis of qualitative variables using more parsimonious quasi-independence models, scaling models, and latent structures", pp. 119-137 in R.K. Merton, J.S. Coleman and P.H. Rossi, eds., Qualitative and quantitative social research. Papers in honor of Paul F. Lazarsfeld. New York: The Free Press.

Goodman, L.A. (1979b). "Simple models for the analysis of association in cross-classifications having ordered categories". Journal of American Statistical Association 74: 537-552.

Goodman, L.A. (1984). The analysis of cross-classified data having ordered categories. Cambridge, Massachusetts: Harvard University Press.

Goodman, L.A. (1987). "New methods for analyzing the intrinsic character of qualitative variables using cross-classified data". American Journal of Sociology 93: 529-583.

Granberg, D. (1978). "Pro-life or reflection of conservative ideology? An analysis of opposition to legalized abortion". Sociology and Social Research 62: 414-429.

Gustaffson, J.E. (1979). PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items. Reports from The Institute of Education, University of Göteborg, Sweden.

Gustaffson, J.E. (1980). "Testing and obtaining fit of data to the Rasch model", British Journal of Mathematical and Statistical Psychology 33: 205-233.

Guttman, L. (1944). "A basis for scaling qualitative data". American Sociological Review 9: 139-150.

Guttman, L. (1950). "The basis for scalogram analysis", pp.60-90 in S. Stouffer, ed., Measurement and Prediction, Princeton, N.J.: Princeton University.

Haberman, S.J. (1974). "Log-linear models for frequency tables with ordered classifications". Biometrics 30: 589-600.

Haertel, E. (1984). "An application of latent class models to assessment data". Applied Psychological Measurement 8: 333-346

Hall, E.J., & Ferree, M.M. (1986). "Race differences in abortion attitudes". Public Opinion Quarterly 50: 193-207.

Hambleton, R.K. (1979). "Latent trait models and their applications" in R. Traub ed., Methodological developments: New directions for testing and Measurement: New directions for testing and measurement. (No. 4). San Francisco: Jossey-Bass

Hambleton, R.K. (1980). "Latent ability scales, interpretations, and uses", in S. Mayo ed., New directions for testing and measurement: Interpreting test scores. (No. 6). San Francisco: Jossey-Bass

Hambleton, R.K., & Cook, L.L. (1977). "Latent trait models and their use in the analysis of educational data". Journal of Educational Measurement 14: 75-96.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory. Principles and applications. Boston: Nijhoff Publishing.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eigner, D.R., & Clifford, J.A. (1978). "Developments in latent trait theory: Models, technical issues and applications". Review of Educational Research 48: 467-510.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory. Boston: Kluwer Nijhoff Publishing.

Hattie, J.A. (1981). Decision criteria for determining unidimensionality. Unpublished doctoral dissertation. University of Toronto, 1981.

Hattie, J. (1984). "An empirical study of various indices for determining unidimensionality", Multivariate Behavioral Research 19: 49-78.

Hattie, J. (1985). "Methodology review: Assessing unidimensionality and tests of items". Applied Psychological Measurement 9: 139-164.

Hayduk, L.A. (1987). Structural equation modeling with LISREL. Essentials and Advances Baltimore: The John Hopkins University Press.

Heise, D.R. (1974). "Some issues in sociological measurement", in H.L. Costner ed., Sociological methodology 1973-1974. London: Jossey-Bass Inc., Publishers.

Hulin, C., Drasgow, F., & Parsons C.K. (1983). Item response theory: application to psychological measurement. Homewood: Dow Jones-Irwin.

Jones, E.F., & C.F. Westhoff (1973). "Changes in attitudes toward abortion: with emphasis upon the national fertility study data", pp.468-481 in H.J. Osofsky and D.J. Osofsky eds., The abortion experience: Psychological and medical impact. New York: Harper and Row.

Kelderman, H. (1984). "Loglinear Rasch model tests". Psychometrika 49: 223-245.

Kingma, J. (1984). "A comparison of four methods of scaling for the acquisition of early number concepts". Journal of General Psychology 110: 23-45.

Kingma, J., & Loth, F.L. (1985). "The validation of a developmental scale for seriation". Educational and Psychological measurement 45: 321-328.

Kingma, J., & Reuvekamp, J. (1984). "The construction of a developmental scale for seriation". Educational and Psychological Measurement 44: 1-23.

Kingma, J., & Ten Vergert, E.M. (1985). "A nonparametric scale analysis of the development of conservation". Applied Psychological Measurement 9: 375-387.

Kingma, J., & Reuvekamp, J. (1986a). "Mokken model: A Pascal program for nonparametric scaling". Educational and Psychological Measurement 46: 667-677.

Kingma, J., & Reuvekamp, J. (1986b). "Mokken test for the robustness of nonparametric scaling". Educational and Psychological Measurement 46: 679-685.

Kingma, J., & Taerum, T. (in press a). "Mokscal: A program for a nonparametric item response model. Applied Psychological Measurement.

Kingma, J., & Taerum, T. (in press b). A program for a nonparametric item response model: The Mokken scale analysis. Psychometrika.

Lazarsfeld, P.F. (1950). "The logical and mathematical foundation of latent structure analysis", pp.362-412 in

S.A. Stouffer and others, Measurement and predictions. Princeton: Princeton University Press.

Lazarsfeld, P.F., & Henry, N.W. (1968). Latent structure analysis. Boston: Houghton Mifflin Company.

Likert, R. (1932). "The method of constructing an attitude scale", pp.233-243 in G.M. Maranell ed., Scaling: A sourcebook for behavioral scientists. Chicago: Aldine Publishing Company.

Loevinger, J. (1948). "The technique for homogeneous tests". Psychological Bulletin 45: 507-529.

Lord, F.M. (1952). A theory of test scores. Psychometrics Monographs. Number 7.

Lord, F.M. (1953). "An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability". Psychometrika 18: 57-76.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale: Lawrence Erlbaum Associates, Publishers.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Lucke, J.F., & Schuessler, K. (1987). "Scaling social life feelings by factor analysis of binary variables", Social Indicator Research 19: 403-428.

Lumsden, J. (1976). "Test theory", Annual Review of Psychology 27: 251-280.

Magnusson, D. (1966). Test theory. London: Addison-Wesley Publishing Comp.

Martin-Löf, P. (1983). Statistiska modeller. Anteckningar från seminarier lasaret 1969-70 utarbetade av Rolf Sundberg. 2:a uppl. (Statistical models. Notes from seminars 1969-1970 by Rolf Sundbaerg. 2nd ed.). Instiutet för försäkringsmatematik och matematisk statistik vid Stockholms Unversiteit.

Masters, G.N., & Wright, B.D. (1981). A model for partial scoring. Research Memorandum Number 31 . Statistical Laboratory Department of Education. The University of Chicago.

McDonald, R.P. (1967). Nonlinear factor analysis. Psychometric Monographs, No.15.

McDonald, R.P. (1981a). "The dimensionality of tests and

items". British Journal of Mathematical and Statistical Psychology 34: 100-117.

McDonald, R.P. (1981b). "Linear versus nonlinear models in item response theory". Applied Psychological Measurement 6: 379-396.

McDonald, R.P. (1982). "Some alternative approaches to the improvement of measurement in education and psychology: Fitting latent trait models, pp.213-237 in D. Spearitt ed., The improvement of measurement in education and psychology. Hawthorn, Australia: Australian Council Educational Residence.

McDonald, R.P. (1985). Factor analysis and related methods. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

McDonald, R.P., & Ahlawat, K.S. (1974). "Difficulty factors in binary data". British Journal of Mathematical and Statistical Psychology 27: 82-99.

McNemar, Q. (1946). "Opinion -attitude methodology". Psychological Bulletin 43: 289-374.

Mead, G.H. (1934). Mind, self, and society. Chicago: University Press.

Mokken, R.J. (1971). A theory and procedure of scale analysis. The Hague: Mouton.

Mokken, R.J., & Lewis, C. (1982). "A nonparametric approach to the analysis of dichotomous item responses". Applied Psychological Measurement 6: 417-430.

Molenaar, W. (1974). "De logistische en de normale kromme" (The logistic and normal curve). Nederlands Tijdschrift voor de Psychologie 29: 415-420.

Molenaar, I.W. (1981). Programmabeschrijving van PML (versie 3.1) voor het Rasch model. (PML user's guide for Rasch model). Rijks Universiteit van Groningen, Vakgroep Statistiek en Meettheorie, Groningen, The Netherlands.

Molenaar, I.W. (1982). "Een tweede weging van de Mokkenschaal" (A second weighing of the Mokken Scale). Tijdschrift voor Onderwijsresearch 7: 172-181.

Molenaar, I.W. (1983). "Rasch, Mokken en schoolbeleving" (Rasch, Mokken and school experience), pp.195-213 in S. Lindenberg and F.N. Stokman eds., Modellen in de sociologie (Models in sociology). Deventer: Van Loghum Slaterus.

Mooijaart, A. (1982). "Latent structure analysis for categorical variables", pp.77-95 in K.G. Jöreskog and H. Wold eds., Systems under indirect observation. Amsterdam: North Holland.

Muthén, B. (1982). "Some categorical response models with continuous latent variables", pp.65-79 in K.G. Joreskog and H. Wold, eds., Contributions to economic analysis. Amsterdam: North-Holland Publishing Co.

Muthén, B. (1983). "Latent variable structural equation modeling with categorical data". Journal of Econometrics 22: 43-65.

Muthén, B. (1984). "A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators". Psychometrika 49: 115-132.

Muthén, B.O. (1987). LISCOMP. Analysis of linear structural equations using a comprehensive measurement model: User's Guide. Mooresville: Scientific Software, Inc.

Nagel, E. (1961). The structure of science. London: Routledge & Kegan Paul.

Neyman, J., & Scott, E.L. (1948). "Consistent estimates based on partially consistent observations". Econometrika 16: 1-32.

Niemöller, B., Verschuur, W., & Stokman, F.R. (1980). Stochastic cummulative scaling: STAP user's manual, Vol 4. Amsterdam, The Netherlands: University of Amsterdam.

Nunally, J.M. (1978). Psychometric theory. New York: McGraw-Hill Book Company.

Petersen, B.L. (1985). Codebook for the Combined 1982 General Social Survey and Allgemeine Bevölkerungsumfrage der Socialwissenschaften . Chicago: University of Chicago.

Proctor, C.H. (1970). "A probabilistic formulation and statistical analysis of Guttman scaling". Psychometrika 35: 73-78.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests . Copenhagen: Denmark paedagogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. Proceedings of the fourth Berkeley Symposium on mathematical statistics and probability 4: 321-333. Berkeley: University of California Press.

Rasch, G. (1966). "An item analysis which takes individual differences into account". The British Journal of Mathematical and Statistical Psychology 19: 49-57.

Reiser, M. (1981). "Latent trait modeling of attitude items", pp.117-144 in G.W. Bohrnstedt and E.F. Borgatta, eds., Social Measurement. Beverly Hills: Sage Publications.

Rindskopf, D. (1983). "A general framework for using latent class analysis to test hierarchical and nonhierarchical learning models". Psychometrika 48: 85-97.

Samejima, F. (1973). "A comment on Birnbaum's three-parameter logistic model in the latent trait theory". Psychometrika 39,: 111-121.

Samejima, F. (1974). "Normal ogive model on the continuous response level in the multi-dimensional latent space". Psychometrika 39: 111-121.

Samejima, F. (1977a). "A use of the information function in tailored testing". Applied Psychological Measurement 1: 233-247.

Samejima, F. (1977b). "Weakly parallel tests in latent trait theory with some criticism of classical test theory". Psychometrika 42: 193-198.

Schuessler, K.F. (1982). Measuring social life feelings. London: Jossey-Bass Publishers.

Schwartz, J.E. (1986). "A general reliability model for categorical data applied to Guttman scales and current-status data, pp.79-119 in N.B. Tuma, ed., Sociological Methodology 1986. San Francisco: Jossey-Bass.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. Tokyo: McGraw-Hill Kogakusha, Ltd.

Shockey, J.W., & Clogg, C.C. (1983). A computer program for the analysis of association in a set of K I-by-J conditional tables. University Park PA: Pennsylvania State University, Population Issues Research Office.

Smith, H.L., & Garnier, M.A. (1987). "Scaling via models for the analysis of association: Social background and educational careers in France", pp. 205-245 in C.C. Clogg, ed., Sociological Methodology 1987. San Francisco: Jossey Bass.

Stevens, S.S. (1951). "Mathematics, measurement, and

psychophysics" in S.S. Stevens ed., Handbook of experimental psychology. New York: John Wiley & Sons Inc.

Stevens, S.S. (1959). "Measurement, psychophysics, and utility", in C.W. Churchman and P. Ratoosh eds., Measurement: definitions and theory. New York: Wiley.

Stinchombe, A.L. (1983). "Linearity in log-linear analysis", pp.104-125 in S. Leinhardt, ed., Sociological methodology 1983-1984. London: Jossey-Bass Publishers.

Stokman, F., & VanSchuur, W. (1980). "Basic scaling". Quality and Quantity 14: 5-30.

Swaminathan, H. (1983). "Parameter estimation in item response models", pp.24-44 in R.K. Hambleton, ed., Applications of Item Response Theory. Vancouver: Educational Research Institute of British Columbia.

Swaminathan, H., & Gkifford, J.A. (1982). "Bayesian estimation in the Rasch model". Journal of Educational Statistics 7: 175-192.

Taylor, G. (1983). "Analyzing qualitative data", pp. 547-612 in P.H. Rossi, J.D. Wright and A.B. Anderson eds., Handbook of survey research. New York: Academic Press, Inc.

Thissen, D.M. (1982). "Marginal maximum likelihood estimation for the one-parameter logistic model". Psychometrika 47: 175-186.

Thissen, D., & Wainer, H. (1982). "Some standard errors in item response theory". Psychometrika 47: 397-412.

Thurstone, L.L. (1927). "A law of comparative judgement". Psychological Review 34 : 273-286.

Torgerson, W.S. (1958). Theory and methods of scaling. New York: John Wiley & Sons Inc.

Van der Linden, W.J. (1978). "Forgetting, guessing, and mastery: The Macready and Dayton models revisited and compared with a latent trait approach". Journal of Educational Statistics 3: 305-318.

Van den Wollenberg, A.L. (1979). The Rasch model and time-limit tests. An application and some theoretical contributions. (Dissertation). Nijmegen: Studentenpers, 1979.

Van den Wollenberg, A.L. (1982a). "Two new test statistics for the Rasch model". Psychometrika 47: 123-140.

Van den Wollenberg, A.L. (1982b). "A simple and effective method to test the dimensionality axiom of the Rasch model". Applied Psychological Measurement 6: 83-91.

Weil, F.D. (1985). "The variable effects of education on liberal attitudes: A comparative-historical analysis of anti-semitism using public opinion survey data". American Sociological Review 50: 458-474.

White, B.W., & Saltz, E. (1957). "Measurement of reproducibility". Psychological Bulletin 54 : 81-99.

Winer, B.J. (1970). Statistical principles in experimental design. London: McGraw-Hill.

Wingersky, M.S. (1983). "LOGIST: A program for computing maximum likelihood procedures for logistic test models, pp. 45-56 in R.K. Hambleton ed., Applications of item response theory. Vancouver: Educational Research Institute of British Columbia.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Wright, B.D., & Masters, G.N. (1981). The measurement of knowledge and attitude. Research Memorandum Number 30. Statistical Laboratory, Department of Education, The University of Chicago.

Wright, B.D., & Panchapakesan, N.A. (1969). "A procedure for sample-free item analysis". Educational and Psychological Measurement 29: 23-48.

Yen, W.M. (1985). "Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory". Psychometrika 50: 399-410.

Yen, W.M. (1986). "The choice of scale for educational measurement: an IRT perspective". Journal of Educational Measurement 4: 299-325.

APPENDIX A

English and German Versions of the Abortion Items

ENGLISH VERSION OF THE QUESTIONS:

Please tell me whether in your opinion a woman should have a legal right to an abortion.....

1. if there is a strong chance of serious defect in the baby.
2. if the woman is married and does not want any more children.
3. if the woman's health is seriously endangered by the pregnancy.
4. if the family has a low income and cannot afford any more children.
5. if the woman became pregnant as a result of rape.
6. if the woman is not married and does not want to marry the father of the child.
7. if the woman wants it for any reason.

GERMAN VERSION OF THE QUESTIONS:

Bitte sagen Sie mir, ob es Ihrer Meinung nach einer Frau gesetzlich möglich sein sollte oder nicht, einen Schwangerschaftsabruch vornehmen zu lassen.....

1. wenn das Baby mit hoher Wahrscheinlichkeit eine ernsthafte Schädigung haben wird.
2. wenn die Frau verheiratet ist und keine Kinder mehr haben möchte.
3. wenn die Gesundheit der Frau durch die Schwangerschaft ernsthaft gefahrdet ist.
4. wenn die Familie nur über ein geringes Einkommen verfügt und sich keine Kinder mehr leisten kann.
5. wenn die Schwangerschaft Folge einer Vergewaltigung ist.
6. wenn die Frau unverheiratet ist und der Vater des Kindes nicht heiraten möchte.
7. wenn die Frau es so will, unabhängig davon, welchen Grund sie dafür hat.

APPENDIX B

Table B.1: The P and P_o Matrix of the Seven Abortion Items for the Americans

Item	A	N	S	P	D	R	H
				Item			
				P-matrix			
A	---	.41	.42	.41	.44	.44	.44
N	.41	---	.46	.47	.50	.51	.51
S	.42	.46	---	.46	.50	.51	.51
P	.41	.47	.46	---	.52	.53	.54
D	.44	.50	.50	.52	---	.81	.84
R	.44	.51	.51	.53	.81	---	.85
H	.44	.51	.51	.54	.84	.85	--
				P_o-matrix			
A	---	.46	.46	.43	.15	.13	.08
N	.46	---	.43	.42	.14	.13	.08
S	.46	.43	---	.41	.14	.13	.08
P	.43	.42	.41	---	.14	.13	.08
D	.15	.14	.14	.14	---	.10	.07
R	.13	.13	.13	.13	.10	---	.07
H	.08	.08	.08	.08	.07	.07	--

Table B.2: The P and P_0 matrix of the Seven Abortion Items for the West-Germans.

Item	A	S	N	P	R	D	H
				Item			
				P-matrix			
A	---	.27	.30	.30	.33	.33	.33
S	.27	---	.30	.31	.34	.34	.34
N	.30	.30	---	.39	.44	.44	.44
P	.30	.31	.39	---	.52	.53	.52
R	.33	.34	.44	.52	---	.88	.88
D	.33	.34	.44	.53	.88	---	.90
H	.33	.34	.44	.52	.88	.90	--
				P_0-matrix			
A	---	.59	.52	.43	.09	.07	.07
S	.59	---	.51	.44	.09	.07	.06
N	.52	.51	---	.42	.09	.07	.06
P	.43	.44	.42	---	.09	.07	.06
D	.09	.09	.09	.09	---	.05	.04
R	.07	.07	.07	.07	.05	---	.04
H	.07	.06	.06	.06	.04	.04	--

APPENDIX C

Table C.1: Latent Class Distributions Under the Models in Table 7.4 for the Americans

UNITED STATES						
Model	Scale Type Probabilities					
	high approval			low approval		
	I	II	III	IV	V	UC
M_1	.50	.01	.03	.32	.14	-
M_2	.43	.01	.02	.24	.14	.16
M_3	.51	.00	.03	.32	.14	-
M_4	.43	.01	.00	.24	.14	.18
M_5	.48	.02	.05	.34	.11	-
M_6	.43	.01	.06	.24	.14	.12
M_7	.49	.02	.03	.32	.14	-
M_8	.42	.01	.01	.25	.15	.16
M_9	.51	.00	.03	.32	.14	-
M_{10}	.44	.01	.00	.23	.13	.18
M_{11}	.40	.00	.02	.24	.13	.21

Note: - refers to parameter not included in given model.

Table C.2: Latent Class Distributions Under the Models in Table 7.4 for West-Germans

	WEST GERMANY					
Model	Scale Type Probabilities					
	high approval				low approval	
	I	II	III	IV	V	UC
M_1	.33	.11	.10	.40	.06	-
M_2	.26	.04	.03	.30	.07	.30
M_3	.33	.10	.10	.40	.07	-
M_4	.34	.09	.01	.17	.00	.28
M_5	.32	.10	.10	.43	.05	-
M_6	.26	.04	.04	.30	.06	.30
M_7	.31	.10	.11	.40	.08	-
M_8	.25	.04	.04	.30	.07	.30
M_9	.33	.11	.11	.38	.06	-
M_{10}	.30	.09	.08	.30	.07	.16
M_{11}	.25	.04	.04	.30	.07	.30

Note: - refers to parameter not included in given model.

Table C.3: Distribution of Guttman Response Patterns According to Ordering (ANY, NOMORE, SINGLE, POOR, DEFECT, RAPE, HEALTH) for the Americans

Response Pattern	Frequency	Percentage
(0,0,0,0,0,0,0)	88	6.82
(0,0,0,0,0,0,1)	31	2.40
(0,0,0,0,0,1,1)	45	3.49
(0,0,0,0,1,1,1)	274	21.24
(0,0,0,1,1,1,1)	40	3.10
(0,0,1,1,1,1,1)	22	1.71
(0,1,1,1,1,1,1)	50	3.88
(1,1,1,1,1,1,1)	500	38.76
Total	1050	81.41
Errors	240	18.59
	1290	100.00

Note: A code of 1 denotes yes; a code of 0 denotes no.

Table C.4: Distribution of Guttman Response Patterns According to Ordering (ANY, SINGLE, NOMORE, POOR, RAPE, DEFECT, HEALTH) for the West Germans

Response Pattern	Frequency	Percentage
(0,0,0,0,0,0,0)	73	3.53
(0,0,0,0,0,0,1)	23	1.11
(0,0,0,0,0,1,1)	62	3.00
(0,0,0,0,1,1,1)	563	27.22
(0,0,0,1,1,1,1)	200	9.67
(0,0,1,1,1,1,1)	153	7.40
(0,1,1,1,1,1,1)	61	2.95
(1,1,1,1,1,1,1)	522	25.24
Total	1657	80.12
Errors	411	19.88
	2068	100.00

Note: A code of 1 denotes yes; a code of 0 denotes no.

Table C.5: Results of the BINO Test of the Seven Abortion Items for Americans in PML

Score group	freq.	positive answers	observed proportion	predicted proportion	One-sided bino
			DEFECT		
1	40	2	.05	.11	.15
2	94	46	.49	.45	.27
3	292	281	.96	.94	.08
4	92	88	.96	.98	.14
5	97	93	.96	.99	.01
6	87	85	.98	.99	.03
Molenaar's U_k=3.19					
			HEALTH		
1	40	31	.78	.72	.28
2	94	86	.92	.91	.51
3	292	290	.99	.99	.49
4	92	91	.99	.99	.27
5	97	96	.99	.99	.13
6	87	85	.98	1.00	.00
Molenaar's U_k=6.65					
			RAPE		
1	40	7	.18	.16	.46
2	94	53	.56	.61	.24
3	292	284	.97	.96	.14
4	92	90	.98	.99	.42
5	97	96	.99	.99	.46
6	87	87	1.00	.99	.84
Molenaars's U_k=-0.25					
			NOMORE		
1	40	0	.00	.00	.93
2	94	0	.00	.00	.45
3	292	5	.02	.03	.17
4	92	18	.20	.27	.07
5	97	58	.60	.55	.19
6	87	78	.90	.83	.05
Molenaar's U_k=-1.91					

Table continued

Score group	freq.	positive answers	observed proportion	predicted proportion	One-sided bino
			POOR		
1	40	0	.00	.00	.89
2	94	2	.02	.01	.34
3	292	9	.03	.04	.19
4	92	46	.50	.41	.04
5	97	66	.68	.69	.49
6	87	72	.83	.89	.06
Molenaar's U_k=1.15					
			SINGLE		
1	40	0	.00	.00	.93
2	94	0	.00	.00	.44
3	292	4	.01	.03	.07
4	92	28	.30	.28	.35
5	97	53	.55	.57	.38
6	87	78	.90	.84	.08
Molenaar's U_k=-1.16					
			ANY		
1	40	0	.00	.00	.98
2	94	1	.01	.00	.22
3	292	3	.01	.00	.48
4	92	7	.08	.09	.46
5	97	23	.24	.22	.34
6	87	37	.43	.45	.35
Molenaar's U_k=0.68.					

Table C.6: Results of the BINO Test for the Seven Abortion Items for the West-Germans in PML

Score group	freq.	positive answers	observed proportion	predicted proportion	One-sided bino
			DEFECT		
1	45	10	.22	.34	.06
2	122	89	.73	.70	.28
3	587	580	.99	.98	.05
4	316	310	.98	.99	.02
5	257	256	.99	.99	.44
6	146	146	1.00	.99	.90
Molenaar's U_k=-.378					
			HEALTH		
1	45	23	.51	.45	.26
2	122	91	.75	.78	.25
3	587	580	.99	.98	.24
4	316	315	.99	.99	.53
5	257	254	.99	.99	.00
6	146	146	1.00	1.00	.92
Molenaar's U_k=1.81					
			RAPE		
1	45	12	.27	.20	.17
2	122	58	.48	.50	.33
3	587	573	.98	.96	.04
4	316	308	.98	.99	.03
5	257	252	.98	.99	.00
6	146	146	1.00	.99	.84
Molenaar's U_k=-2.11					
			NOMORE		
1	45	0	.00	.00	.91
2	122	4	.03	.00	.00
3	587	7	.01	.02	.10
4	316	64	.20	.27	.00
5	257	192	.75	.68	.01
6	146	134	.92	.88	.12
Molenaar's U_k=-0.22					

Table continued

Score group	freq.	positive answers	observed proportion	predicted proportion	One-sided bino
			POOR		
1	45	0	.00	.01	.81
2	122	1	.00	.02	.46
3	587	13	.02	.05	.00
4	316	208	.66	.60	.01
5	257	218	.85	.85	.51
6	146	133	.91	.95	.04
Molenaar's U_k=0.45					
			SINGLE		
1	45	0	.00	.00	.97
2	122	1	.01	.00	.22
3	587	3	.01	.01	.52
4	316	28	.09	.08	.39
5	257	68	.27	.26	.42
6	146	86	.59	.61	.30
Molenaar's U_k=.89					
			ANY		
1	45	0	.00	.00	.98
2	122	0	.00	.00	.81
3	587	5	.00	.00	.21
4	316	31	.10	.07	.06
5	257	45	.18	.23	.03
6	146	85	.58	.56	.30
Molenaar's U_k=0.38					

Table C.7: Results of the BINO Test for the Five Abortion Items for Americans in PML

Score group	freq.	positive answers	observed proportion	predicted proportion	One-sided bino
			RAPE		
1	343	334	.97	.97	.23
2	98	95	.97	.99	.13
3	98	96	.98	.99	.09
4	88	96	.99	.99	.87
Molenaars's U_k=1.25					
			NOMORE		
1	343	2	.00	.00	.40
2	98	19	.19	.26	.09
3	98	55	.56	.55	.42
4	88	79	.90	.83	.06
Molenaar's U_k=-1.83					
			POOR		
1	393	4	.01	.01	.46
2	98	47	.48	.41	.08
3	98	68	.69	.69	.53
4	88	72	.82	.89	.02
Molenaar's U_k=1.58					
			SINGLE		
1	343	0	.00	.00	.04
2	98	28	.29	.27	.41
3	98	53	.54	.56	.36
4	88	78	.89	.84	.15
Molenaar's U_k=-0.52					
			ANY		
1	343	3	.00	.00	.07
2	98	7	.07	.08	.50
3	98	22	.22	.20	.34
4	88	35	.40	.44	.26
Molenaar's U_k=0.17					

Table C.8: Results from the Fischer and Schleibechner Test of the Five Abortion Items for the American high/low Scoregroups

Pair-wise T-Tests of the equality of item parameters			
Variable	Item Parameters low	high	Difference high-low
Rape	-3.97	-2.96	-1.69
Nomore	-1.10	- .41	2.74
Poor	.21	.27	-0.29
Single	.81	.48	1.42
Any	1.85	1.80	0.16

Table C.9: Results of the BINO Test for the Four Abortion Items for West-Germans in PML

Score group	freq.	positive answers	observed proportion	predicted proportion	One-sided bino
			DEFECT		
1	134	82	.61	.63	.36
2	1017	1012	.99	.99	.19
3	268	267	.99	.99	.41
Molenaars's U_k=-.51					
			RAPE		
1	134	51	.38	.36	.38
2	1017	1004	.99	.99	.48
3	268	264	.99	.99	.03
Molenaar's U_k=-1.94					
			SINGLE		
1	134	1	.00	.00	.30
2	1017	9	.00	.01	.27
3	268	146	.55	.54	.43
Molenaar's U_k=.92					
			ANY		
1	134	0	.00	.00	.72
2	1017	9	.00	.01	.44
3	268	127	.47	.47	.45
Molenaar's U_k=-0.22					

Table C.10: Results from the Fischer and Schleibechner Test of the Four Abortion Items for the West-German high/low Scoregroups

Pair-wise T-Tests of the equality of item parameters			
Variable	Item Parameters low	high	Difference high-low
DEFECT	-2.97	-2.78	-0.26
RAPE	-2.45	-1.40	-2.71
SINGLE	2.66	2.02	3.62
ANY	2.76	2.16	3.33

Table C.11: Item and Person Parameter Estimates of the Four Abortion Items for the West Germans

Variable	Difficulty δ	Person Parameter θ
DEFECT	-2.89	-2.64 (sum score 1)
RAPE	-2.35	.02 (sum score 2)
SINGLE	2.55	2.65 (sum score 3)
ANY	2.69	

Table C.12: Person Parameter, Item Parameter, and the Rasch Probability of a Positive Answer of the Items RAPE and POOR for the Americans

Person Parameter θ	Item Parameter RAPE Item b_k	Difference $(\theta - b_k)$	Probability
-2.29	-3.88	1.59	.83
- .19	-3.88	3.69	.97
.96	-3.88	4.84	.99
2.16	-3.88	6.04	1.00

Person Parameter θ	Item Parameter POOR ITEM b_k	Difference $(\theta - b_k)$	Probability
-2.29	.34	-2.63	.07
- .19	.34	- .53	.37
.96	.34	.62	.65
2.16	.34	1.82	.86

Table C.13: LOGIST Item Parameters and Standard Errors of the Two Parameter Model of the Seven Abortion Items for the Americans.

	Two-Parameter Model			
	a	std.err.	δ	std.err.
DEFECT	1.16	.2	-1.38	.2
HEALTH	.39	.4	-5.47	31.2
RAPE	1.72	.2	-1.41	.1
NOMORE	2.49	.2	.91	.1
POOR	1.60	.2	.76	.1
SINGLE	2.48	.2	.89	.1
ANY	1.06	.2	1.76	.2

Table C.14: LOGIST and NOHARM Item Parameter Estimates of the Two Parameter Model of the Seven Abortion Items for the Americans (N=702) and the West-Germans (N=1473)

	AMERICANS			
	LOGIST		NOHARM	
	a	δ	a	δ
DEFECT	1.16	-1.38	.52	-2.21
HEALTH	.39	-5.47	.11	-16.19
RAPE	1.72	-1.41	1.02	-1.64
NOMORE	2.49	.91	2.10	.83
POOR	1.60	.76	1.06	.81
SINGLE	2.48	.89	1.34	.91
ANY	1.06	1.76	.54	2.69
	WEST-GERMANS			
	LOGIST		NOHARM	
	a	δ	a	δ
DEFECT	1.36	-2.09	.95	-2.30
HEALTH	1.09	-2.44	.56	-3.50
RAPE	1.15	-1.90	.58	-2.74
NOMORE	2.37	.77	1.84	.70
POOR	2.12	.40	1.05	.39
SINGLE	1.31	1.47	.59	2.26
ANY	1.20	1.59	.52	2.62

Table C.15: NOHARM Item Parameter Estimates of the One Parameter Model of the Seven Abortion Items For the Americans

	N=1290	N=702
	Item Parameter δ	Item Parameter δ
DEFECT	-1.20	-1.19
HEALTH	-1.58	-2.14
RAPE	-1.29	-1.36
NOMORE	- .03	.87
POOR	- .11	.68
SINGLE	- .04	.85
ANY	.17	1.48

FIGURE C8.1: Model with dichotomous variables and ABORT as reference variable

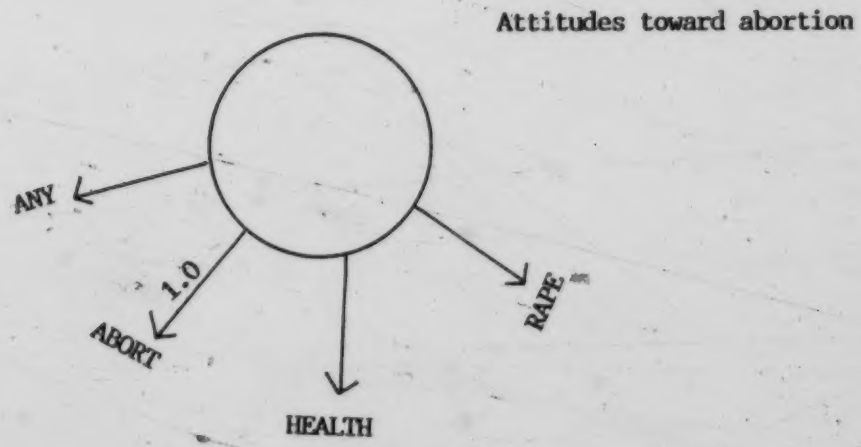
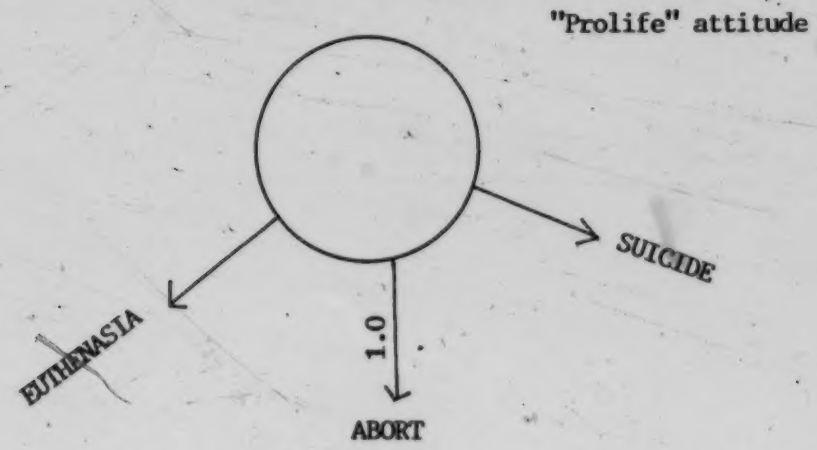



FIGURE C8.2: Model with continuous variables and ABORT as reference variable